Census 2000 Experiment

April 17, 2003

Administrative Records Experiment in 2000 (AREX 2000) Household Level Analysis

FINAL REPORT

This research paper reports the results of research and analysis undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation (TXE) Program, designed to assess Census 2000 and to inform 2010 Census planning. Findings from the Census 2000 TXE Program reports are integrated into topic reports that provide context and background for broader interpretation of results.

Mark Bauder and
D. H. Judson

Planning, Research, and
Evaluation Division

U S C E N S U S B U R E A U

Helping You Make Informed Decisions

Intentionally Blank

ACKNOWLEDGMENTS

The Administrative Records Experiment 2000 was conducted by the staff of Administrative Records Research at the U.S. Census Bureau, led by Charlene Leggieri. Questions and comments regarding this document can be directed to Mark Bauder or Dean Judson at 301-457-4229 / 4222.

Administrative Records Research Staff Members and Key Contributors to AREX 2000:

Bashir Ahmed	Mikhail Batkhan	Mark Bauder
Mike Berning	Harold Bobbitt	Barry Bye
Benita Dawson	Joseph Conklin	Kathy Conklin
Gary Chappell	Ralph Cook	Ann Daniele
Matt Falkenstein	Eleni Franklin	James Farber
Mark Gorsak	Harley Heimovitz	Fred Holloman
David Hilnbrand	Dave Hubble	Robert Jeffrey
Dean Judson	Norman Kaplan	Vickie Kee
Francina Kerr	Jeong Kim	Myoung Ouk Kim
Charlene Leggieri	John Long	John Lukasiewicz
Mark Moran	Daniella Mungo	Esther Miller
Tamany Mulder	Nancy Osbourn	Arona Pistiner
Ron Prevost	Dean Resnick	Pamela Ricks
Paul Riley	Douglas Sater	Doug Scheffler
Kevin A. Shaw	Kevin M. Shaw	Larry Sink
Diane Simmons	Amy Symens-Smith	Cotty Smith
Herbert Thompson	Deborah Wagner	Phyllis Walton
Signe Wetrogan	David Word	Mary Untch

and

Members of the AREX 2000 Implementation Group

Intentionally Blank

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Intentionally Blank

EXECUTIVE SUMMARY

In the Administrative Records Experiment in 2000 (AREX 2000), an administrative records census was conducted in which administrative records were used to enumerate people and obtain demographic data. The Administrative Records Experiment was conducted in two sites: one composed of two counties in Maryland, and the other composed of three counties in Colorado. The Administrative Records Experiment 2000 was intended to compare methodologies for conducting an administrative records census, and to evaluate the results of this administrative records census.

Two methodologies for conducting an administrative records census were tested in Administrative Records Experiment 2000. This evaluation focuses on one of these: the Bottom-up method. In the Bottom-up method, administrative records persons are grouped into households, and administrative records addresses are linked with addresses in an independently maintained address list.

The primary goals of this evaluation were to assess the coverage and accuracy of household level data from the administrative records census, and to investigate the feasibility of using administrative records to substitute for nonresponse in a survey or census.

We assessed the coverage and accuracy of Administrative Records Experiment by comparing its results to those of Census 2000. In order to investigate the feasibility of using administrative records to substitute for nonresponse, Administrative Records Experiment data were compared to Census data for Census non-responding households. Our analyses considered two kinds of household level Census nonresponse: Nonresponse Followup households, and "imputed households" (or "imputed housing units"). Imputed households include: those whose vacancy status is unknown after mailout/mailback and Nonresponse Followup operations have been completed; and those which are known to be occupied, but contain no data defined people after mailout/mailback and Nonresponse Followup operations have been completed.

Key findings of the evaluation include the following:

- **Coverage of the Census universe**. Administrative Records Experiment housing units could be linked with:
    - about 81 percent of Census housing units and 84 percent of occupied Census housing units,
    - about 71 percent of Nonresponse Followup housing units and 77 percent of occupied Nonresponse Followup housing units, and
    - about 62 percent of Census imputed housing units, and 63 percent of imputed housing units that were imputed to be occupied.
- **Comparison of household size**. Among matched, occupied housing units, Administrative Records Experiment and Census household sizes were the same for:
    - about 51 percent of all households,
    - about 37 percent of Census Nonresponse Followup households,
    - about 32 percent of imputed households imputed to be occupied, and
    - about 27 percent of imputed households imputed to be vacant.

- **Comparison of household demographic composition**. Among linked households of the same size, Administrative Records Experiment and Census agreed in demographic composition (age, sex, Hispanic origin, and age groups 0-17, 18-64, 85+) in:
    - about 81 percent of all households,
    - about 63 percent of Nonresponse Followup households, and
    - about 23 percent of imputed households.
- **Households which were covered less well by the Administrative Records Experiment, or had more discrepancy between Census and the Administrative Records Experiment for size or demographic composition**. We found several types of households for which administrative records did less well with regard to coverage or accuracy. These include:
    - Households within multi-unit structures. Census households that were within a multi-unit structure were less likely to be linked with Administrative Records Experiment households. When such households were linked, Administrative Records Experiment and Census were less likely to agree in household size and household demographic composition.
    - Households containing races other than White, or Hispanics. For Census households that contained people of races other than White, or contained Hispanics, Administrative Records Experiment and Census agreed less often in size and demographic composition than for other households.
    - Households in which a race was imputed in the Administrative Records Experiment. Administrative Records Experiment and Census distributions of racial composition were more similar for Administrative Records Experiment households in which no person's race was imputed, than when all Administrative Records Experiment households were included.
- **Predicting households in which Administrative Records Experiment household characteristics agreed well with Census characteristics**. We developed a model which predicted, with 72.1 percent accuracy, when an Administrative Records Experiment household's demographic composition (size, and the fully crossed array of: sex, race, Hispanic origin, and age in 5-year categories) was the same as the linked Census household's demographic composition. We found some characteristics of Administrative Records Experiment households that were useful predictors of Administrative Records Experiment and Census demographic equivalence. These include:
    - being in a single unit structure,
    - containing only one or two persons,
    - containing no persons with imputed race,
    - containing one or more White persons, and
    - containing only persons 65 and older in the household.

We also found substantial interaction effects in the model:

- Administrative Records Experiment households which are in single unit structures, contain only persons 65 and older, and have no imputed race, are five times more likely to match in demographic composition than other households.
- Administrative Records Experiment households in which there are one or more White persons, only one or two persons, and only persons are 65 and older, are 19 times more likely to match in demographic composition than other households.

On the basis of the results of this evaluation, we recommend the following:

- **Improve record linkage techniques**. The success of a Bottom-Up style administrative records census depends on the ability to link addresses. While 80 percent of Census households were linked with Administrative Records Experiment households, the percentage of Nonresponse Followup and imputed households that were linked was significantly lower. Research should continue into new computer methods for linkage of records, and for parsing and standardizing addresses. Clerical review processes should be should be used to resolve many-to-one and one-to-many address links.
- **Investigate ways to reduce the time lag between administrative records and surveys or censuses**. The time lag between administrative records used in the Administrative Records Experiment and Census date appears to be a major reason for discrepancies between the Administrative Records Experiment and Census results. Ways to reduce the time lag between administrative records and when they are available for nonresponse substitution should be investigated. In particular, the possibility of obtaining and processing records on a flow basis should be investigated.
- **Improve race and Hispanic origin imputation**. Imputation of race and Hispanic origin were a source of inaccuracies of Administrative Records Experiment demographic data. Research should continue into the development of models to impute race and Hispanic origin.
- **Continue to explore techniques for predicting when administrative records household level data are likely to be accurate**. Suppose that the accuracy of administrative records has not been proven accurate enough for nonresponse substitution in a particular survey or census. Administrative records might still be accurate enough to substitute for some types of non-responding households in that survey or census. Modeling techniques should be developed to predict households at which administrative records are likely to be accurate.
- **Test the use of administrative records for substitution for nonresponse**. With the lessons learned in the Administrative Records Experiment, improved methods for conducting an administrative records census can be developed. These improved methods should be tested. Future Census tests would be ideal candidates for these tests. These tests could evaluate the accuracy and coverage of administrative records data, the quality of record linkage operations, and the validity of models used to predict households for which administrative records are particularly accurate. Finally, tests could be done in which proposals for nonresponse substitution are implemented.

Intentionally Blank

1. BACKGROUND

1.1 Introduction

The Administrative Records Experiment 2000 (AREX 2000) was an experiment in two areas of the country designed to gain information regarding the feasibility of conducting an administrative records census (ARC), or the use of administrative records in support of conventional decennial census processes. The first experiment of its kind, AREX 2000 was part of the Census 2000 Testing, Experimentation, and Evaluation Program. The focus of this program was to measure the effectiveness of new techniques, methodologies, and technologies for decennial census enumeration. The results of the testing lead to formulating recommendations for subsequent testing and ultimately to the design of the next decennial census.

Interest in taking a decennial census by administrative records dates back at least as far as a proposal by Alvey and Scheuren (1982) wherein records from the Internal Revenue Service (IRS) along with those of several other agencies might form the core of an administrative record census. Knott (1991) identifies two basic ARC models: (1) the Top-down model that assembles administrative records from a number of sources, unduplicates them, assigns geographic codes and counts the results; and (2) the Bottom-up model that matches administrative records to a master address file, fills the addresses with individuals, resolves gaps and inconsistencies address by address, and counts the results. There have been a number of other calls for ARC research — see for example Myrskyla 1991; Myrskyla, Taeuber and Knott 1996; Czajka, Moreno and Shirm 1997; Bye 1997. All of the proposals fit either the Top-down or Bottom-up model described here.

Knott also suggested a composite Top-down/Bottom-up model, which would unduplicate administrative records using the Social Security Number (SSN) then match the address file and proceed as in the Bottom-up approach. In overall concept, AREX 2000 most closely resembles this composite approach.

More recently, direct use of administrative records in support of decennial applications was cited in several proposals during the Census 2000 debates on sampling for Nonresponse Followup (NRFU). The proposals ranged from direct substitution of administrative data for non-responding households (Zanutto, 1996; Zanutto and Zaslavsky, 1996; 1997; 2001), to augmenting the Master Address File development process with U.S. Postal Service address lists (Edmonston and Schultze, 1995:103). AREX 2000 provided the opportunity to explore the possibility of NRFU support.

The Administrative Records Research (ARR) staff of the Planning, Research, and Evaluation Division (PRED) performed the majority of coordination, design, file handling, and certain field operations of the experiment. Various other divisions within the Census Bureau, including Field Division, Decennial Systems and Contracts Management Office, Population Division, and Geography Division supported the ARR staff.

Throughout this report, rather than identifying individual workgroups or teams, we shall refer to the operational decisions made in support of AREX to be those of ARR; that is, we shall say that "ARR decided to…" whenever a key operational decision is described, even though, of course, ARR staff were not the only decision makers.

1.2 Administrative Record Census—Definition and Requirements

In the AREX, an administrative record census was defined as a process that relies primarily, but not necessarily exclusively, on administrative records to produce the population content of the decennial census short form with a strong focus on apportionment and redistricting requirements. Title 13, United States Code, directs the Census Bureau to provide state population counts to the President for the apportionment of Congressional seats within nine months of Census Day. In addition to total population counts by state, the decennial census must provide counts of the voting age population (18 and over) by race and Hispanic origin for small geographic areas, currently in the form of Census blocks, as prescribed by PL 94-171 (1975) and the Voting Rights Act (1964). These data are used to construct and evaluate state and local legislative districts.

Demographically, the AREX provided date of birth, race, Hispanic origin, and sex, although the latter is not required for apportionment or redistricting purposes. Geographically, the AREX operated at the level of basic street address and corresponding Census block code. Unit numbers for multi-unit dwellings were used in certain address matching operations and one of the evaluations; but generally, household and family composition were not captured. In addition, the design did not provide for the collection of sample long form population or housing data, needs that will presumably be met in the future by the American Community Survey program. The design did assume the existence of a Master Address File and geographic coding capability similar to that available for the Census 2000.

1.3 AREX Objectives

The principal objectives of AREX 2000 were twofold. The first objective was to develop and compare two methods for conducting an administrative records census, one that used only administrative records and a second that added some conventional support to the process in order to complete the enumeration. The evaluation of the results also included a comparison to Census 2000 results in the experimental sites.

The second objective was to test the potential use of administrative records data for some part of the NRFU universe, or in place of other imputation methods. In order to effectively use administrative records databases for substitution purposes; one must determine which kinds of administrative record households are most likely to yield similar demographic distributions to their corresponding census households.

Other more general objectives of the AREX included the collection of relevant information, available only in 2000, to support ongoing research and planning for administrative records use in the 2010 Census, and the comparison of an administrative records census to other potential

2010 methodologies. These evaluations and other data will provide assistance in planning major components of future decennial censuses, particularly those that have administrative records as their primary source of data.

1.4 AREX Top-down and Bottom-up Methods

1.4.1 Top-down

The AREX 2000 enumeration was accomplished by a two-phase process. The first phase involved the assembly and computer geocoding of records from a number of national administrative record systems, and unduplication of individuals within the combined systems. This was followed by two attempts to obtain and code physical addresses (clerical geocoding and request for physical address) for those that would not geocode by computer. Finally, there is a selection of "best" demographic characteristics for each individual and "best" street address within the experimental sites. Much of the computer processing for this phase was performed as part of the Statistical Administrative Records System (StARS) 1999 processing (Judson, 1999; Farber and Leggieri, 2002). As such, StARS 1999 was an integral part of the AREX 2000 design.

One can think about the results of the Top-down process in two ways. First, counting the population at this point provides, in effect, an administrative-records-only census. That is, the enumeration includes only those individuals found in the administrative records, and there is no other support for the census outside of activities related to geocoding. AREX 2000 provides population counts from the Top-down phase so that the efficacy of an administrative-records-only census can be assessed.

However, without a national population register as its base, one might expect an enumeration that used only administrative records to be substantially incomplete. Therefore, a second way to think about the Top-down process is as a substitute for an initial mail-out in the context of a more conventional census that would include additional support for the enumeration.

1.4.2 Bottom-up

The fundamental difference between the Bottom-up method and the Top-down method is the Bottom-up method matches administrative records addresses to a separately developed "frame" of addresses, and based on this match, performs additional operations. In this experiment, an extract of the Census Bureau's Master Address File (MAF) served as the frame[1].

The second phase of the AREX 2000 design was an attempt to complete the administrative-records-only enumeration by the correction of errors in administrative records addresses through address verification (a coverage improvement analogue) and by adding persons missed in the administrative records (a NRFU analogue). This phase began by matching the addresses found in the Top-down process to the MAF in order to assess their validity and to identify those MAF addresses for which no administrative records were found. A field address review (FAV) was used to verify non-matched administrative records addresses, and invalid administrative records addresses were excluded from the Bottom-up selection of best address. Non-matched MAF addresses were canvassed in order to enumerate persons at addresses not found in the administrative records systems. In the AREX, such a canvassing was simulated by adding those persons found in the Census 2000 at the unmatched addresses to the adjusted administrative-records-only counts, thus completing the enumeration. Accomplishing the AREX as part of the

[1] In this report, we use the term "MAF" generically. Our operations were based on extracts from the Decennial Master Address File (DMAF).

Census 2000 obviated the need to mount a separate field operation to canvass unmatched MAF addresses.

Considering the Top-down and Bottom-up processes as part of one overall design, AREX can be thought of as a prototype for a more or less conventional census with the initial mailout replaced by a Top-down administrative records enumeration. Figure 1 below, provides a conceptual overview of the experiment for enumerating the population tested during the AREX. A more detailed description of data processing flows can be found in Attachment 1.

Note: The graphical description presented here is intended to convey the concept of both AREX methods when viewed in terms of the Bottom-up method as a follow-on process to the Top-down method.

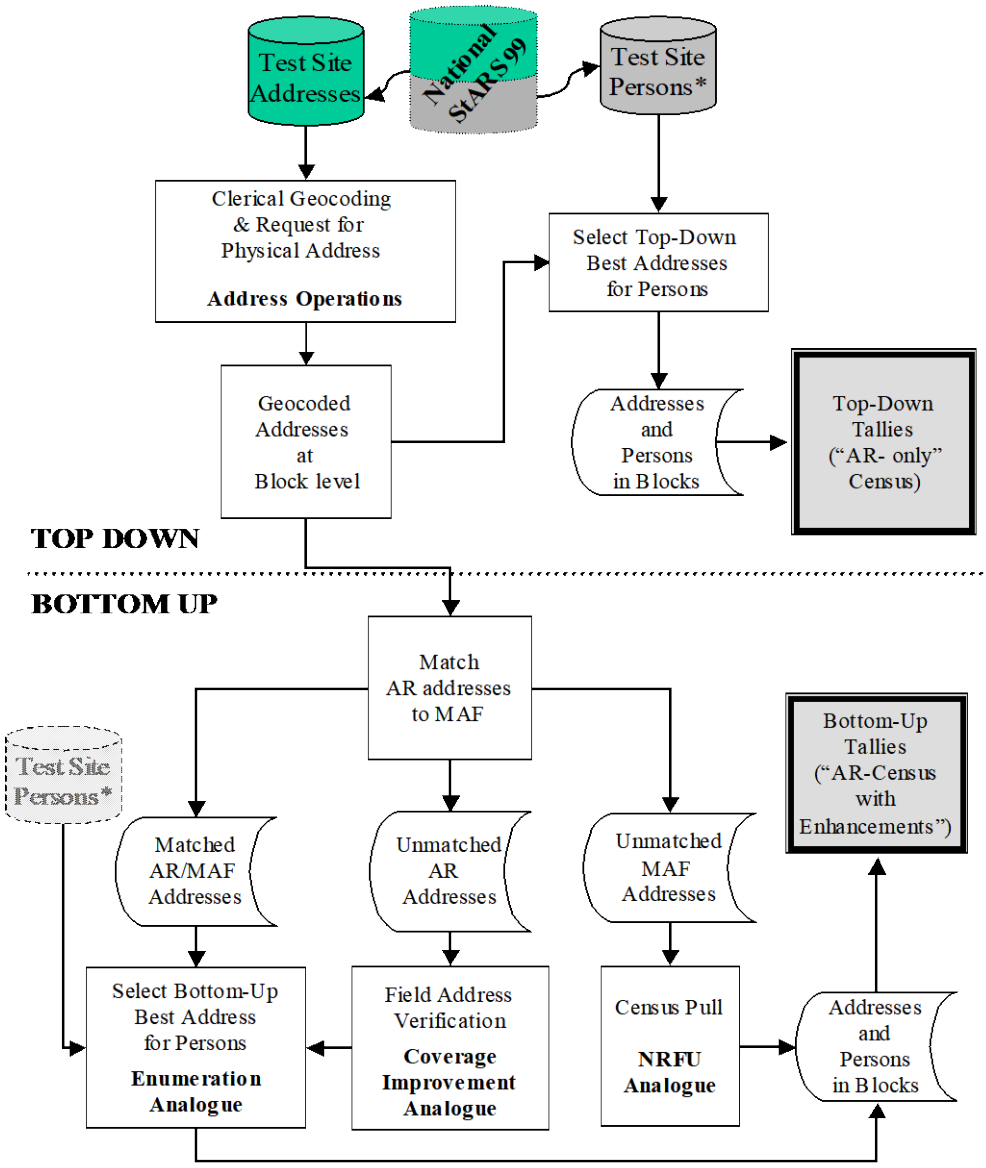


Figure 1. Summary Diagram of AREX 2000 Design

1.5 Experimental Sites

The experiment was set up to include geographic areas that include both difficult and easy to enumerate populations. Two sites were selected believed to have approximately one million housing units and a population of approximately two million persons. One site included Baltimore City and Baltimore County, Maryland. The other site included Douglas, El Paso, and Jefferson Counties, Colorado. The sites provided a mix of characteristics needed to assess the difficulties that might arise in conducting an administrative records census. Approximately one half of the test housing units was selected based on criteria assumed to be easy-to-capture in an administrative records census (for example, areas having a preponderance of city style addresses, single family housing units, older and less mobile populations), and the other half was selected based on criteria assumed to be hard to capture (the converse). Demographic characteristics of the sites are given in the following table.

Table 1. Key Demographic Characteristics of the AREX 2000 Sites

	Baltimore County, MD	Baltimore City, MD	Douglas County, CO	El Paso County, CO	Jefferson County, CO	United States
Total Population[1]	754,292	651,154	175,766	516,929	527,056	281,421,906
White[1]	74.4%	31.6%	92.8%	81.2%	90.6%	75.1%
Black[1]	20.1%	64.3%	1.0%	6.5%	0.9%	12.3%
American Indian, Eskimo or Aleut[1]	0.3%	0.3%	0.4%	0.9%	0.8%	0.9%
Asian or Pacific Islander[1]	3.2%	1.5%	2.6%	2.7%	2.4%	3.7%
Other Race[1]	0.6%	0.7%	1.4%	4.7%	3.2%	5.5%
Multi-Race[1]	1.4%	1.5%	1.9%	3.9%	2.2%	2.4%
Hispanic[1]	1.8%	1.7%	5.1%	11.3%	10.0%	12.5%
Median Age[1]	37.7 yrs	35.0 yrs	33.7 yrs	33.0 yrs	36.8 yrs	35.3 yrs
Crude Birth Rate[2]	12.6	14.9	19.0	15.7	12.5	14.9[3]
Crude Death Rate[2]	9.9	13.1	2.7	5.5	6.0	8.6[3]
1990-2000 Change[4]	9.0%	-11.5%	191.0%	30.2%	20.2%	13.2%

Note: All values include household and group quarters residents.
[1] 2000 Census results
[2] 1998 rates per 1000; from MD Dept. of Health and Mental Hygiene and CO Dept. of Public Health and Environment
[3] 1998 rates per 1000; from www.fedstats.gov
[4] 1990 and 2000 Census results

1.6 AREX Source Files

The administrative records for AREX were drawn from the StARS 1999 database. There were six national-level source files selected for inclusion in StARS. The files were chosen to provide the broadest coverage possible of the U.S. population, and to compensate for the weaknesses or lack of coverage of a given segment of the population inherent in any one source file. The national level files that contributed to the StARS 1999 database and to AREX 2000 included the following:

- Internal Revenue Service (IRS) Tax Year 1998 Individual Master File (1040),
- IRS Tax Year 1998 Information Returns File (W-2 / 1099),
- Department of Housing and Urban Development (HUD) 1999 Tenant Rental Assistance Certification System (TRACS) File,
- Center for Medicare and Medicaid Services (CMS) 1999 Medicare Enrollment Database (MEDB) File,
- Indian Health Services (IHS) 1999 Patient Registration System File, and
- Selective Service System (SSS) 1999 Registration File.

The following table displays the primary reason each file was included in the StARS database and the approximate number of input records associated with each.

Table 2. Source File Characteristics

File	Targeted Population Segment	Address Records	Person Records
IRS 1040	Taxpayer and other members of the reporting unit) with current address	120 million	243 million
IRS W2/1099	Persons with taxable income who might not have filed tax returns	598 million	556 million
HUD TRACS	Low income housing population (possible non-taxpayers)	3 million	3 million
Medicare File	Elderly population (possible non-taxpayers)	57 million	57 million
IHS File	Native American population (possible non-taxpayers)	3 million	3 million
SSS File	Young male population (possible non-taxpayers)	14 million	13 million
	Total	**795 million**	**875 million**

Notes: Variance between the number of address records and person records within input source files is a result of the following source file characteristic anomalies.

1. The number of address records column is generally synonymous with the total record count on the input file.
2. Each IRS 1040 input record may reflect up to six persons (primary filer, secondary filer, and four dependents).
3. Each SSS input record may reflect two addresses - defined as current and/or permanent address.
4. The IRS W-2/1099 file undergoes a preliminary unduplication and clean-up process prior to the initial file edit process. Prior to person processing, records are written "out of scope" if the SSN field is blank, the edited or input name field is blank, or the name standardizer returns a "bad name" — such as institutional or firm names.

1.6.1 Timing

An important limitation for the AREX is the gap between the reference period for data contained in each source file and the point-in-time reference of April 1, 2000 for the Census. The time lag has an impact on both population coverage—births, deaths, immigration and emigration—and geographic location—housing extant, and geographic mobility. As an example, both IRS files include data for tax year 1998 with an expected current address as of tax filing time close to April 15, 1999. Note, however, that the IRS 1040 file only provided persons in the tax unit as of December 31, 1998. The following table provides the reference periods of the files available. Generally, the reference periods are about one year prior to the 2000 Census day.

Table 3. Currency of Source Files

Source File	Cut-off Date	Requested Cut Date	Universe
Indian Health Svc.	04/01/99	04/01/99	All persons alive at cut-off date
Selective Service	Note 2	04/01/99	Males between the age of 18 - 25[2]
HUD TRACS	04/01/99	04/01/99	All persons on file as of cut-off date
Medicare	Note 3	04/01/99	All persons alive at cut-off date
IRS 1040	12/98	09/30/99[1]	Individual tax returns for tax year 1998
IRS W-2 / 1099	12/98	04/01/99[1]	Forms W-2 and 1099 forms for tax year 1998

1. File Cut date is for posting cycle weeks 1-39 only for IRS 1040, and weeks 1-41 for IRS 1099 files. Weeks 40-52 (and 42-52 respectively) were not included in StARS '99.

2. Cut-off date is same as dates used to define universe: persons born after April 2, 1972 and on (or before) April 1, 1980.

3. Universe also defined as persons with a death date of 12/31/1989 or later.

1.6.2 State, Local and Commercial Files

ARR staff decided not to use state and local files[2] and commercially available databases[3] in the AREX 2000 experiment. Statistical evidence is limited, but various reports from ARR staff indicated that state and local files come in an extremely diverse variety of forms, with equally diverse record layouts and content (for historical information, see Sweet, 1997; Buser, Huang, Kim, and Marquis, 1998; and other papers in the Administrative Records Memorandum Series). Furthermore, ARR staff reported that it was quite time-consuming and intricate to develop the interagency contractual arrangements necessary to use state and local files. Public opinion results such as Singer and Miller (1992), Aguirre International (1995), and Gellman (1997), convinced ARR staff that public sensitivity to the idea of linking commercial databases with government databases (other than for address processing) would be too great, and that such a linkage would be unwise.

In addition to acquisition and processing difficulties, consideration of the use of state and local files raises an equity issue in a decennial census context. Since it is not possible to obtain an exact count of the population in its entirety, public perception of fair treatment in the decennial census process is important. Therefore, the accuracy of the counts must be seen as uniform between and within states. The use of data from only certain states or localities would compromise notions that decennial census methods must treat all parts of the country equitably.

The American Business Index (or ABI) file was used to identify addresses that were commercial rather than residential, and a Group One product, Code One, used to standardize addresses.

[2] Such as state and local tax returns, drivers license files, local utilities, assessor's records, and the like.
[3] Such as commercially available mailing lists, credit card databases, and the like.

1.6.3 Census Numident

An additional, and critical, file used in creation of the StARS database was the Census Numident file. For the AREX, it was the source of most of the demographic characteristics and some of the death data.

The Census Numident was created by ARR for the primary purpose of validating Social Security Numbers (SSNs) used in the processing of administrative records and supplying demographic variables missing from source files. The Census Numident is an edited version of the Social Security Administration's (SSA) Numerical Identification (Numident) File. The SSA Numident file is the numerically ordered master file of assigned Social Security Numbers (SSN) that may contain up to 300 entries for each SSN record, although on average contains two records per SSN. Each entry represents an initial application for a SSN or an addition or change (referred to as a transaction) to the information pertaining to a given SSN. The SSA Numident contains all transactions (and therefore, multiple entries) ever recorded against a single SSN. The SSA Numident available for StARS 1999 reflected all transactions through December 1998.

The Census Numident was designed to collapse the SSA Numident entries to reflect "one best record" for each SSN containing the "best" demographic data for each SSN on the file. However, all variations in name data (including married names, maiden names, nicknames, etc.) and all variations in date of birth data were retained as part of the Census Numident as an Alternate Name File and Alternate Date of Birth File, respectively. For the Census Numident, selection criteria were established for each of the desired Census 2000 Short Form demographic variables (after minor edits were accomplished in an effort to standardize the variables). The short form variables included such items as date of birth, gender, race, and Hispanic origin. Following edit, unduplication, and selection processing, the SSA Numident file of nearly 677 million records was reduced to just over 396 million records that comprise the Census Numident file.

1.7 AREX Evaluations

Currently, four evaluations are being completed.

The **Process Evaluation** documents and analyzes selected components or processes of the Top-down and Bottom-up methods in order to identify errors or deficiencies. It is designed to catalog the various processes by which raw administrative data became final AREX counts and attempts to identify the relative contributions of these various processes.

The **Outcomes Evaluation** is a comparison of Top-down and Bottom-up AREX counts by county, tract, and block level counts of the total population by race, Hispanic origin, age groups and gender, with comparable decennial census counts. This evaluation is outcome rather than process oriented.

The **Household Evaluation** assesses outcomes of the Bottom-up method, the potential for NRFU substitution and household imputations, and predictive capability. NRFU substitution assesses the feasibility of using administrative records, in lieu of a field interview, to obtain data on non-responding census addresses via the Bottom-up method.

The **Request for Physical Address Evaluation** assesses the impact of noncity-style addresses. These addresses present a significant hurdle to the use of an administrative records census on either a supplemental or substitution basis is the determination of residential addresses and their

associated geographic block level allocation for individuals whose administrative record address is a P.O. Box or Rural Route. AREX 2000 tested a possible solution in the form of the Request for Physical Address operation. Several thousand letters were mailed to P.O. Box and Rural Route addresses requesting the receiver to reply with their residential address for purposes of block level geocoding. This report documents in detail the planning and implementation of the operation. It also analyzes the results of the operation and assesses its potential future use as part of an ARC.

2. METHODOLOGY

2.1 What were the general goals of the household evaluation?

The general goal of this evaluation is to focus on household-level comparisons. In the process, we will examine several difficult to measure aspects of the enumeration process: Nonresponse Followup (NRFU) households, and households for which occupancy status, household population, and/or household demographics were wholly imputed ("imputed households"). We will specifically assess the ability of AREX databases to match the demographic distributions of all households, NRFU households, and imputed households. Finally, we will attempt to assess our ability to predict *when* an AREX household is likely to demographically match a census household.

2.1.1 NRFU evaluation

Addresses with missing enumeration forms must be investigated by Nonresponse Followup procedures (NRFU). NRFU addresses are the most expensive to enumerate and may represent the most vulnerable segment of Americans. The evaluation considers whether AREX can replace or reduce more expensive NRFU processing by examining NRFU addresses, their socio-demographic characteristics, and how these vary at high and low levels of geography.

2.1.2 Imputed households evaluation

Here, we use the term "imputed households" to refer to households for which occupancy status, population count, or all demographics were imputed for Census 2000. The evaluation considers the evidence that AREX databases offer on the occupancy and demographic characteristics of imputed households.

2.1.3 Prediction

One of the most important potential uses of administrative records data is to *substitute* administrative records data for some proportion of the Nonresponse Followup universe, or for the imputed households[4]. In order to effectively use administrative records databases for substitution purposes, we must determine which kinds of administrative record households are most likely to yield similar demographic distributions to their corresponding census households. The purpose of the prediction section is to make this evaluation.

[4] A related use is to use administrative records data to improve non-interview weighting for nonresponse in surveys; this also requires matching and substitution, but will not be considered here.

2.2 What special terminology do we use in this evaluation?

We use the term *"census household"* when referring to an address populated by persons with a relationship to the householder. AREX processing connects persons with addresses, but no relationship to a householder is determined. Therefore, we use the term *"AREX household"* for the people at one address, with the understanding that "relationship to householder" information is not contained in the AREX database. For convenience, we apply this definition to vacant housing units, so that when a housing unit contains no people, we will consider it to contain a household of size zero. *"Household size"* refers to the number of people in the housing unit. Group Quarters are excluded for the analyses here, so all addresses we consider are addresses of housing units.

We refer to a pair of addresses (AREX and Census) that were linked through a computerized record linkage process as *"linked"* housing units. We use the term *"linked households"* when comparing the properties of people within linked housing units. We use the term *"imputed household"* or "*whole household imputation*" for households for which occupancy status, population count, and/or all household demographic characteristics have been imputed. We use the term *"demographic match"* when two households have the same distribution of age, race, sex and Hispanic origin.

Finally, we will use the term *"AREX data"* for data obtained from the BARCUF file (BARCUF stands for "Bottom-Up Administrative Records Census Unedited File"), the resulting file from simulated Bottom-Up operations. We will use the term *"Census data"* for data obtained from the HDF file (HDF stands for "Hundred Percent Detail File"). In this analysis, we did not use the "Census Pull" addresses that were analyzed in the outcomes and process evaluations. These AREX addresses, because they were taken from the HDF file, by construction contain the same people.

2.3 What were the fundamental dependent variables?

The fundamental dependent variables in the modeling phases of this evaluation are comparisons between two distributions, that of the decennial census and that of AREX, at the (computer linked) address level. There are two distributions of main interest, the age/sex distribution and the race/ethnicity distribution. The measure we chose to model asks: Do the addresses match on the fully crossed distributions (that is, the age distribution *by* the sex distribution *by* the race distribution *by* the ethnicity distribution)? This measure is represented by an indicator variable:

$$\text{Match} = \begin{cases} 1 & \text{if the fully crossed age} \times \text{race} \times \text{sex} \times \text{Hispanic origin distributions in the} \\ & \text{linked Census household match the AREX household;} \\ 0 & \text{otherwise.} \end{cases}$$

This measure is based on the *distribution* of personal characteristics within an address. Thus, it has a substantial weakness: If an address in Census that is matched to a address in AREX that has similar demographic characteristics, *but is composed of entirely different persons*, the match indicator could still indicate agreement. While this is not problematic distributionally, it is problematic from an enumeration point of view[5].

[5] We would like to note that there is a second, more stringent measure of success we proposed as our fundamental dependent variable: Matched persons in matched addresses. This most stringent dependent variable would simply

2.4 What descriptive analyses do we perform?

We perform descriptive analyses for the full five county AREX universe, for the Census NRFU universe, and for imputed households. In these analyses, we compare household level characteristics of AREX and Census. In particular, we do the following:

- Evaluate the coverage by AREX of its intended universe by determining the number and proportion of Census addresses that were matched by AREX addresses;
- Examine the effect of properties of Census households on the proportion of Census households that were matched;
- Compare AREX and Census distributions of household size and household demographic characteristics for the AREX universe and subsets of the AREX universe;
- Compare household size and demographic characteristics of AREX and Census matched households; and
- Examine the effect of household properties on the comparisons of distributions, and on household to household comparisons. Examples of such household properties include: the presence of a person in the household of a particular race or ethnicity, and the presence of a person with a characteristic that was imputed in AREX.

2.5 How do we know when an AREX address will be similar to a census address?

A final part of the evaluation will consist of attempting to model the situations where we can predict that an AREX address will have similar demographic characteristics to a Census address.

2.6 Applying Quality Assurance Procedures

Quality assurance procedures were applied to the design, implementation, analysis, and preparation of this report. The procedures encompassed methodology, specification of project procedures and software, computer system design and review, development of clerical and computer procedures, and data analysis and report writing. A description of the procedures used is provided in the "Census 2000 Evaluation Program Quality Assurance Process."

3. LIMITS

3.1 Operational limits that limit the scope of this evaluation

3.1.1 Group Quarters

Because of operational limitations Group Quarters' counts were eliminated on the AREX 2000 database (for those persons for whom ARRS determined that their address was a Group Quarter or Special Place). In order to make block counts and distributions comparable, persons enumerated in a Group Quarter or Special Place in Census 2000 were also eliminated. In an actual administrative records census ARRS would field an actual Group Quarters operation, most likely similar to existing Group Quarters and Special Place enumerations. For the purposes of the AREX 2000 simulation, this field operation was not conducted.Administrative Records

---

be an indicator that all persons within the linked addresses could themselves be linked. This measure of matching was not implemented for this AREX evaluation.

In an administrative records census using the bottom up method, addresses that were in the DMAF address file but not identified in the AREX database would be enumerated in an "administrative records Nonresponse Followup" operations. These operations would likely consist of some combination of mailout/mailback, telephone and/or field operations. These follow-up operations could not be supported for AREX 2000. For the AREX experiment, Census data were included in AREX for unmatched Census addresses. In this way, Census mail-out and NRFU operations were used as substitutes for those that would have been done in a true administrative records census.

3.1.2 Field Address Verification

A feature of the original design of the AREX 2000 experiment was the inclusion of a coverage improvement simulation. A field address verification operation was to be performed on 100 percent of the AREX 2000 addresses that did not match to the DMAF. However, because of Census 2000 requirements, that verification could not be performed. Rather than omit field address verification information, a sample-based operation was performed. This information has been incorporated into the bottom-up method. However, despite the use of a sample in the AREX 2000 experiment, it should be recognized that ARRS strongly prefers a 100 percent field address verification operation, rather than a sample.

3.1.3 Other File Limitations

Several individual limitations of the files themselves are worthy of note: First, AREX 2000 used files that were a year or more older than the target date of Census day. This means that movers, births, deaths, immigration and emigration, new housing, abandoned and demolished housing are unaccounted for. Second, AREX 2000 by definition has difficulty enumerating children properly, by virtue of the time lag problem and by virtue of the limited demographics available for children on the Numident file (Miller, Judson, and Sater, 2000). Third, the race measurement and reporting deficiencies of the AREX 2000 experiment cause comparisons by race and Hispanic origin to be more challenging. In particular, most persons of Hispanic origin were imputed as such by AREX, thus complicating comparisons. Of course, Census 2000 multiple race reporting additionally complicates comparisons between AREX and Census households.

3.2 General limitations

The major limitation of this study is that it is observational in nature rather than experimental. The characteristics used as regressors in the model developed in section four are not controlled by the researcher but rather are random variables. Consequently the tabulations and modeling are primarily descriptive and the hypothesis tests used to determine any coefficient effects are not strictly correct. They should be understood as guidelines for future model building.

A second major limitation of this study is that the sample of blocks in the five counties in which the AREX experiment was performed are neither statistically representative of Census 2000 blocks nor some superpopulation of blocks. Because of this, we cannot make proper statistical inferences about AREX/Census 2000 relationships in either 2000 or in the hypothetical superpopulation. Therefore, any inferential results presented should be considered as guidelines to future model building and identified as approximate.

3.3 Limitations on the interpretation of the results

In some of the analyses below, we compared results for households in the Census NRFU universe with other households. The distinction we used between non-NRFU and NRFU households does not exactly correspond with the distinction between households for which the Census data is from mailout/mailback returns and other households. Households could have been characterized as in NRFU, yet ultimately a mail return was used to provide the data for the household. And some households were not characterized as NRFU at the time the NRFU universe was set, but ultimately enumerator data (rather than a mail return) was the source of Census data for the household.

In addition, the Census file we used as our "reference" file for determining whether AREX demographics matched Census demographics was the 100% Detail File, or HDF. The tabulations from the HDF match those that are publicly available, for example from American Factfinder. However, due to confidentiality constraints on release of data, some confidentiality protections have been imposed on the HDF file, in particular "swapping" of individuals from hous ehold to household. In situations where such "swapping" occurred, the AREX demographics may match the "unswapped" households, but of course not match the two "swapped" households. The "correct" match status would not be reflected in our results.

4. RESULTS

4.1 Descriptive Results

The purposes of the Descriptive Results section of this report are to compare AREX household level data to Census household level data, and determine how these comparisons vary with characteristics of the household. For most of the evaluations here, we treat Census as truth. Thus, when evaluating how the AREX to Census comparisons vary with the true household characteristics, we use characteristics of the *Census* household. By contrast, in Section 4.2 below, the purpose is to show how we might *predict* households in which administrative records data are correct. For that purpose, the household size and demographic variables used as predictors are *AREX* variables. Only these would be available for a non-responding household, if we were truly trying to use administrative records to substitute for nonresponse.

4.1.1 Why do we use the results of the Bottom-Up method for the analyses here?

In the Bottom-Up method, administrative records addresses are matched to an independently maintained address list. This address list might not be the same as the lists used for surveys or a Census. If these administrative records were to be used to substitute for nonresponse in a survey or a census, an address match would be required. The addresses in the administrative records address list would have to be linked with the survey or census addresses. In the particular implementation of the Bottom-Up method *for AREX*, the Census Bureau's MAF was used as the independently maintained address file. Since the MAF contains Census addresses, the address match between AREX and Census has already been done, and we can use the links between AREX addresses and the MAF in our comparisons of AREX and Census linked addresses.

4.1.2 What are the basic household level characteristics of the AREX and Census Universes?

- *What administrative records data and Census data are compared in this evaluation?*

We compare the results of the AREX Bottom-Up method to the Census Bureau's Hundred Percent Detail File (HDF), which is the source for Census 2000 data that were released and available on, for example, the American Factfinder. In the analyses in this evaluation, 'AREX' refers to the operations and results of the Bottom-Up method, and 'Census' to HDF.

- *What are the basic characteristics of Census address data?*

In the five counties covered by the AREX experiment, Census contains 1,092,460 housing units (HUs) and 1,744 group quarters (GQs). Because AREX contains no administrative records data for Census GQs, we do not include Census GQs in later analyses. There are 24,584 "imputed households[6]" in Census, accounting for 2.3 percent of all Census households. The Census NRFU universe contains 360,914 households, which is 33.0 percent of the total.

[6] Recall that we adopt the convention that a vacant housing unit contains a household of size zero.

- *What are the basic characteristics of AREX address data?*

As part of the implementation of the Bottom-Up method for AREX, Census data were included in the AREX results for Census addresses with which no administrative records could be linked. We do not include them in the analyses, because we want to analyze the coverage and accuracy of *administrative records* data. There are 1,065,031 remaining AREX addresses.

Of these 1,065,031 AREX addresses, 56,638 were not linked with any DMAF address, and1,008,393 were linked with DMAF addresses. The version of the DMAF that was used in the matching process was earlier than that used for Census 2000. Thus, not all of the DMAF addresses available for the AREX matching process still existed in Census 2000. Of the linked AREX addresses, 992,865 were linked with addresses that exist in Census. Of those that were linked with Census addresses, 889,638 are "perfect matches." These are linked AREX—Census address pairs in which each address was linked with exactly one address. There were "non-perfect matches" – both where an AREX address was linked with more than one DMAF address, and where more than one AREX address was linked with one DMAF address. With further processing, we may have been able to resolve some of the "non-perfect" matches. However, we believe that the number of them is small enough that those statistical analyses that use linked addresses will not be affected substantially. In what follows, "linked" addresses are always those that were perfect matches.

Some AREX addresses are flagged as GQs, based on DMAF records for the linked addresses. As noted above, no administrative records data were used for the 1,744 GQs in Census. However, 128 of the AREX addresses that were flagged as GQs remained in AREX. Of these, 90 were linked with Census *housing units* (not GQs). Of the 90, 61 had perfect matches to Census housing units. For analyses of linked addresses below, the 61 perfect matches to Census HUs are included. However, for other analyses, all 128 AREX GQs are left out of the analysis – to be consistent with the fact that we leave all 1,744 Census GQs out of the analyses.

4.1.3 How well did AREX cover the Census universe?

In this evaluation, we intend to evaluate the ability of administrative records to substitute for or supplement a census. Thus, when we speak of the "coverage" by AREX of a Census universe, we are referring to the number or proportion of Census housing units with which we could associate AREX administrative records data.

- *How well did AREX cover the universe of Census addresses, for occupied and vacant addresses?*

Of the 1,092,460 Census housing unit addresses, 889,638 (81.4 percent) were linked with AREX addresses. Because the administrative records files used for AREX typically contain only *person* records, we expected that AREX would not cover vacant addresses as well as occupied ones. The data confirm this expectation. AREX housing units were linked with 84.0 percent of the 1,017,273 occupied Census housing units. AREX housing units were linked with 46.4 percent of the 75,187 vacant Census housing units. We give more detailed information in Table 4 below.

Table 4. Coverage by AREX of Census Housing Units

	Total	Linked with AREX housing units (% of Total)	Linked with AREX occupied housing units (% of Total)	Linked with AREX vacant housing units (% of Total)
Census housing units	1,092,460	889,638 (81.4%)	813,688 (74.5%)	75,950 (7.0%)
Occupied Census housing units	1,017,273	854,741 (84.0%)	787,802 (77.4%)	66,939 (6.6%)
Vacant Census housing units	75,187	34,897 (46.4%)	25,886 (34.4%)	9,011 (12.0%)

- *How well did AREX cover the universe of Census NRFU housing units and of Census imputed households?*

AREX did not cover the Census NRFU universe as fully as it did the non-NRFU universe. AREX housing units were linked with 70.9 percent of the 360,914 Census NRFU housing units, compared with 86.6 percent of the Census non-NRFU housing units. For occupied NRFU housing units, the coverage rate goes up to 76.7 percent. Table 5 contains more details about AREX coverage of Census NRFU and non-NRFU housing units.

Table 5. Coverage by AREX of Census Housing Units by NRFU Status

Type of Census housing unit	Total	Linked with AREX housing units	Linked with AREX occupied housing units	Linked with AREX vacant housing units
NRFU	360,914	70.9%	60.8%	10.1%
Non-NRFU	731,546	86.6%	81.2%	5.4%
Occupied NRFU	289,224	76.7%	67.1%	9.6%
Occupied non-NRFU	728,049	86.9%	81.5%	5.4%
Vacant NRFU	71,690	47.6%	35.2%	12.3%
Vacant non-NRFU	3,497	22.4%	17.7%	4.7%

There are 24,584 imputed housing units in Census. AREX housing units were linked with 62.3 percent of them. AREX addresses were linked with 63.2 percent of those that were imputed to have people in them, and 34.7 percent of those imputed to be vacant. We give more details in Table 6.

Table 6. Coverage by AREX of Census housing units, by imputation status

Type of Census housing unit	Total	Linked with AREX housing units	Linked with occupied AREX housing units	Linked with vacant AREX housing units
Imputed	24,584	62.3%	51.7%	10.5%
Non-imputed	1,067,876	81.9%	75.0%	6.9%
Imputed occupied	23,811	63.2%	52.6%	10.6%
Non-imputed, occupied	993,462	84.5%	78.0%	6.5%
Imputed vacant	773	34.7%	25.5%	9.2%
Non-imputed, vacant	74,414	46.5%	34.5%	12.0%

We see that AREX addresses linked with a smaller proportion of NRFU housing units than of non-NRFU housing units; and linked with a smaller proportion of imputed housing units than non-imputed housing units. The under coverage of NRFU and imputed households can be due to several factors. Among them are the following:

- Address data from NRFU and/or imputed housing units might be generally of lower quality, and thus harder to match.
- Addresses of these housing units may be of types that are harder to match, e.g., those in apartment buildings, those on Rural Routes, or at P.O. boxes.
- People in these housing units may be less likely to have records in any of the administrative records used for AREX.

In addition, households that were imputed to be occupied in Census may easily have been vacant. In that case, we would not expect to have administrative records from the housing unit.

- *How did coverage vary for subsets of the NRFU universe?*

Within the NRFU universe, some of the households were more difficult to get data from. These are cases where it would especially attractive to use administrative records. In Table 7, we consider two subgroups of NRFU: imputed households (as before, but here only those in NRFU), and those households where enumerators got data from a proxy (here, either someone at the address who did not live there in Census date, or a neighbor, etc.).

Table 7. Coverage by AREX of Census housing units by type of NRFU household. Occupied Census housing units only

Type of Census housing unit	Total	Linked with AREX housing units
Non-NRFU	728,049	632,832 (86.9%)
NRFU, not imputed or proxy	228,354	179,961 (78.8%)
NRFU, proxy response	39,779	27,919 (70.2%)
NRFU, imputed	21,091	14,029 (66.5%)

We see that coverage for these more difficult NRFU cases was somewhat worse, but not much. Coverage was about 67 percent for the imputed households, 70 percent for proxy cases, compared to about 79 percent for the rest of NRFU.

4.1.4 How do the sizes of AREX and Census households compare?

- *How do the distributions of household size compare between AREX and Census?*

We use the term "household size" to refer to the number of people in the household, *i.e*., the number of people in the housing unit. We will adopt the convention that a vacant housing unit contains a household of size zero. For many of the analyses, we do not include vacant housing units, because we know that the AREX covers them much less well. Table 8 shows the distributions of household size for AREX and for Census. Tables B.1 through B.7 in Appendix B contain more detailed comparisons for the AREX universe, and for each of the five counties.

The AREX distribution of household size is nearly identical to the Census distribution. We consider it promising that these distributions are so similar. One small pattern that we can see is that AREX almost always has a smaller percentage of two person households. From tables B.1 through B.7 in Appendix B, we can see that this is true for each of the five counties.

Table 8. Distributions of household size for Census and AREX for all five AREX counties — occupied housing units only

Household Size	Census		AREX	
	Total	%[1]	Total	%[2]
1	276,590	27.2%	246,726	27.9%
2	331,472	32.6%	262,075	29.6%
3	171,136	16.8%	155,929	17.6%
4	142,822	14.0%	127,295	14.4%
5	60,988	6.0%	56,596	6.4%
6	21,655	2.1%	22,695	2.6%
7-9	11,275	1.1%	12,481	1.4%
10+	1,335	0.1%	1,625	0.2%
All Sizes	1,017,273	100%	885,422	100%

[1] Percent of all Census occupied housing units

[2] Percent of all AREX occupied housing units

From Tables B.1 through B.7 in Appendix B, we note that among the *unlinked* housing units in both Census and AREX, a very high percentage have one person according the respective file. One possible explanation of this fact is that a much higher percentage of one-person households are at basic street addresses[7] (BSAs) at which there are multiple housing units, and addresses at such BSAs are harder to match. We test this hypothesis by comparing match rates by Census household size and by whether the Census address is at a multi-unit BSA. The results are in Table 9.

We see that, conditional on whether a household is at a multi-unit BSA, the match rates are nearly constant across size of the household. Ignoring vacant addresses, coverage rates for Census housing units at multi-unit BSAs are consistently at about 67 percent, while coverage rates at single-unit BSAs are consistently at about 90 percent. We conclude that whether an address is at a multi-unit BSA has a significant effect on whether an AREX housing unit was linked with the housing unit. We also conclude that, once the difference between multi- and single units is taken into account, the size of the household has little effect on coverage rates.

[7] Two addresses are at the same BSA if they are identical except for apartment numbers or other unit identifiers.

Housing units at multi-unit BSAs are harder to match because matching requires agreement on apartment number or other unit identifier. Among possible reasons that such addresses are harder to match are:

- unit identifiers are often written in different forms,
- unit identifiers are sometimes left off, and
- unit identifiers probably entered less accurately than other address fields.

Table 9. Coverage of Census housing units by Census size and by multi-unit vs. single-unit

Census HH Size	All HUs		Multi-Unit		Single-Unit	
	Total	Linked*	Total	Linked	Total	Linked
All Sizes	1,092,460	81.4%	312,363	64.4%	780,097	88.3%
0	75,187	46.4%	33,916	36.4%	41,271	54.7%
1 or more	1,017,273	84.0%	278,447	67.8%	738,826	90.1%
1	276,590	78.3%	135,833	67.0%	140,757	89.2%
2	331,472	85.2%	80,719	69.2%	250,753	90.4%
3	171,136	86.2%	33,162	68.7%	137,974	90.4%
4	142,822	87.8%	18,082	68.4%	124,740	90.6%
5	60,988	87.1%	6,992	64.7%	53,996	90.0%
6	21,655	86.7%	2,398	64.2%	19,257	89.5%
7+	12,610	86.6%	1,261	57.7%	11,349	89.8%

* Linked with an AREX housing unit via an address match.

4.1.5 How do the distributions of AREX and Census household characteristics compare?

The analyses below concern all Census housing units, and AREX housing units. No GQs in either file are included in the analyses.

- *How do the distributions of demographic characteristics of households compare between AREX and Census?*

Table 6 contains information about the distributions of household level race characteristics for AREX and Census. Occupied housing units are characterized by whether they:

- contain only Whites,
- contain only people of the same race, but not White, or
- contain people of more than one race.

We compare Census and AREX according to their distributions in the above categories. We believe that to best evaluate the accuracy of *administrative records* data, we should compare those AREX households within which no person's AREX race was imputed. Thus, our tables contain distributions both for those AREX households with no imputed race, and for all the AREX households. Table 6 contains the distribution of household race characteristics for the

full AREX universe. In Tables B.7 through B.11 in Appendix B, we compare distributions in more detail, and for each of the five counties.

In general, the AREX distribution of household race characteristics for those with no imputed race is similar to that of Census. We can see one pattern. Compared with Census, AREX generally has a slightly higher percentage of all-White households and a slightly lower percentage of households composed entirely of one race which is not White. This pattern holds for households containing four or fewer people. The same pattern occurs for each of the counties, with one exception. In Baltimore City, which, according to Census, has a much higher percentage of Blacks than the other counties, the percentages of households with all of one race other than White are much more similar to those of Census.

We can also note that, when we include AREX households including those with imputed races, the AREX distribution is generally not as close to the Census one. In particular, AREX tends to have more mixed race households. The effect of race imputations in AREX is discussed in a later section of the paper.

Table 10. Distributions of household race characteristics for Census and AREX households

HH Size		Households with all Whites		Households with all one race other than White		Mixed race Households		Totals	
		# of HHs	(%)[1]	# of HHs	(%)[1]	# of HHs	(%)[2]	Total[2]	(%)
	Census	205,139	(74.2%)	71,451	(25.8%)	N/A		**276,590**	**(100%)**
1	AREX (No imputed race)	178,739	(76.1%)	56,297	(24.0%)	N/A		**235,036**	**(100%)**
	AREX (total)	187,763	(76.3%)	58,477	(23.7%)	N/A		**246,240**	**(100%)**
	Census	256,496	(77.4%)	62,452	(18.8%)	12,524	(3.8%)	**331,472**	**(100%)**
2	AREX (No imputed race)	185,228	(80.5%)	37,460	(16.3%)	7,461	(3.2%)	**230,149**	**(100%)**
	AREX (total)	204,965	(78.6%)	44,378	(17.0%)	11,376	(4.4%)	**260,719**	**(100%)**
	Census	116,767	(68.2%)	45,108	(26.4%)	9,261	(5.4%)	**171,136**	**(100%)**
3	AREX (No imputed race)	68,047	(70.0%)	24,712	(25.4%)	4,462	(4.6%)	**97,221**	**(100%)**
	AREX (total)	107,474	(69.6%)	36,588	(23.7%)	10,256	(6.6%)	**154,318**	**(100%)**
	Census	102,127	(71.5%)	32,600	(22.8%)	8,085	(5.7%)	**142,822**	**(100%)**
4	AREX (No imputed race)	42,592	(71.9%)	13,740	(23.2%)	2,869	(4.9%)	**59,138**	**(100%)**
	AREX (total)	92,029	(73.1%)	24,399	(19.4%)	9,531	(7.6%)	**125,959**	**(100%)**
	Census	40,412	(66.3%)	16,674	(27.3%)	3,902	(6.4%)	**60,988**	**(100%)**
5	AREX (No imputed race)	13,568	(63.8%)	6,277	(29.5%)	1,412	(6.6%)	**21,257**	**(100%)**
	AREX (total)	37,533	(67.2%)	13,024	(23.3%)	5,323	(9.5%)	**55,880**	**(100%)**
	Census	12,700	(58.6%)	7,269	(33.6%)	1,686	(7.8%)	**21,655**	**(100%)**
6	AREX (No imputed race)	3,867	(54.0%)	2,639	(36.8%)	657	(9.2%)	**7,163**	**(100%)**
	AREX (total)	13,127	(58.6%)	6,533	(29.2%)	2,723	(12.2%)	**22,383**	**(100%)**
	Census	5,990	(47.5%)	5,390	(42.7%)	1,230	(9.8%)	**12,610**	**(100%)**
7+	AREX (No imputed race)	1,136	(33.5%)	1,806	(53.5%)	434	(12.9%)	**3,376**	**(100%)**
	AREX (total)	5,427	(39.0%)	6,011	(43.3%)	2,460	(17.7%)	**13,898**	**(100%)**

[1]Percent of Total

[2] Households with no people whose race was missing

4.1.6 How do the sizes of linked housing units compare?

There are 889,638 perfect match address pairs, representing 81.4 percent of the Census housing units and 83.5 percent of AREX housing units. These linked pairs are used in the analyses below.

- *How do household sizes compare between AREX and Census? How do they compare for NRFU housing units, and for whole household imputations?*

Comparisons of household size for linked AREX-Census housing units are given in Table 11. Here is a summary of the results.

AREX and Census counted the same number of people in the housing unit (*i.e., in the household)* for 51.1 percent of the 889,638 linked households. For 79.4 percent of the linked housing units, the AREX person count was within one of the person Census count.

AREX counted the same number as Census for 56.8 percent of the linked Census non-NRFU housing units. For linked NRFU housing units, the AREX count was the same as the Census count for 37.0 percent. The AREX count is within one of the Census count for 83.5 percent of non-NRFU housing units, and was within one of the Census count for 69.3 percent of the NRFU housing units.

We saw above that AREX did not cover the NRFU universe as well as it did other Census housing units. Among Census households with which AREX households are linked, AREX had the same number of people as Census for a smaller percent of NRFU households than other households. This could be because:

- more people move out/move in for NRFU households, or
- administrative records are less accurate or complete for the types of people that tend to be in NRFU households, or
- *Census* data are less accurate for NRFU households.

AREX had the same count for 51.4 percent of the 874,327 linked non-imputed Census housing units, and was within one of the Census count for 79.6 percent. For the 15,043 linked imputed occupied households, AREX had the same count for 31.8 percent, and was within one for 66.8 percent of these addresses. For the 268 linked imputed vacant housing units, AREX also had a count of zero for 26.5 percent, and had a count of zero or one for 62.0 percent.

The low percentage of household by household agreement for imputed households between AREX and Census household should be expected. Since these are imputations on the *Census* side, the best that could be hoped for Census is that the distribution over some larger population of households is correct. The comparison of AREX and Census for Census imputed housing units is a test of the imputation method more than of the accuracy of AREX.

Table 11. Comparison of Census and AREX household size, by NRFU status, and by imputation status—for linked housing units

AREX person count compared with Census	All Census housing units	Census non-NRFU housing units	Census NRFU housing Units	Non-imputed Census housing units	Imputed vacant Census housing units	Imputed occupied Census housing units
Same count	454,437 (51.1%)*	359818 (56.8%)	94,619 (37.0%)	449,582 (51.4%)	71 (26.5%)	4,784 (31.8%)
AREX one higher than	124,706 (14.0%)	84,269 (13.3%)	40,437 (15.8%)	122,519 (14.0%)	95 (35.5%)	2,092 (13.9%)
AREX one lower	127,531 (14.3%)	85,178 (13.4%)	42,353 (16.5%)	124,355 (14.2%)	0	3,176 (21.1%)
AREX 2 or 3 higher	64,635 (7.3%)	36769 (5.8%)	27,866 (10.9%)	63,024 (7.2%)	77 (28.7%)	1,534 (10.2%)
AREX 2 or 3 lower	79,848 (9.0%)	47,938 (7.6%)	31,910 (12.5%)	77,463 (8.9%)	0	2,385 (15.9%)
AREX 4 or more higher	15,781 (1.8%)	6,486 (1.0%)	9,295 (3.6%)	15,316 (1.8%)	25 (9.3%)	440 (2.9%)
AREX 4 or more lower	22,700 (2.6%)	13,158 (2.1%)	9,542 (3.7%)	22,068 (2.5%)	0	632 (4.2%)
Total	**889,638 (100%)**	**633,616 (100%)**	**256,022 (100%)**	**874,327 (100%)**	**268 (100%)**	**15,043 (100%)**

* Percents are percents of column total

In Charts B.12 and Charts B.13 in Appendix B ,we plot the distributions of AREX household sizes for fixed Census household sizes, and the distributions of Census household sizes for fixed AREX household sizes. These distributions are discussed below.

Ignoring distributions for households of size zero, for each Census size up through six, the mode of the distribution of AREX household size is the Census size. Above Census size of six, the AREX mode remains at six. Thus AREX consistently undercounts large Census households. Note also that, for *AREX* households of size greater than six, the mode of the distribution of Census household size is six or fewer.

We've seen that for large Census households, and for large AREX households, the Census household size tends to be smaller. These tendencies may represent a "regression toward the mean." We know that, because of the time lag between our administrative records and Census, there will sometimes be different people at an address in AREX than in Census. In such a case, when there were many people in AREX at the address, we would expect that the household that moved into the housing unit later would be smaller. Similarly, where there were many people in the Census household at the address, there usually would have been fewer AREX people in the address before the Census people moved in. A test of whether the time lag between our administrative records and Census accounts for this phenomenon has not been done.

There is another reason to expect that when the Census household size is greater than six, the AREX household size is most likely to be six. The largest source of administrative records used in AREX is the IRS 1040 file. IRS provides the Census Bureau up to four dependents per tax return. Thus we expect that, when the household size was greater than six, we still did not get

records for more than six people. Hence AREX would tend to undercount Census households of sizes greater than six.

- *How do household size comparisons vary for different kinds of NRFU households?*

As with coverage, for household size comparisons, we considered also some of the "difficult" NRFU cases: imputed households, and those where data came from a proxy.

Table 12. Household size comparisons for subsets of the NRFU Universe. Linked, occupied Census housing units only

Type of Census housing unit	Total	AREX and Census Have Equal Household Size	AREX within one of Census Size
Non-NRFU	632,832	359,652 (56.8%)	528,769 (83.6%)
NRFU, not imputed or proxy	179,961	70,837 (39.4%)	128,270 (71.3%)
NRFU, proxy response	27,919	10,672 (38.2%)	20,642 (73.9%)
NRFU, imputed	14,029	4,265 (30.4%)	9,271 (66.1%)

For these comparisons, there was virtually no difference between imputed households, or proxy cases, and the rest of NRFU.

4.1.7 How do the demographic properties of linked households compare?

In the next few analyses, we compare demographic characteristics of linked households. Because comparisons within households of different sizes are difficult to interpret, we consider only linked occupied housing units in which AREX and Census have the same number of people. There are 445,426 of these housing units representing 40.8 percent of all Census housing units, 41.8 percent of all AREX housing units, and 51.2 percent of all linked housing units.

- *How often do AREX and Census agree about numbers in basic demographic categories? How do these comparisons differ between the NRFU and non-NRFU universes?*

Table 8 contains data only for linked households for which AREX and Census had the same total count. The table shows the frequencies with which AREX and Census agree for:

- each sex category;
- each race category: White, Black, American Indian, Asian/Pacific Islander;
- each Hispanic origin category;

- each five-year age category: 0-4, 5-9, …, 80-84, 85 and up;
- and each of the age categories: 0-17, 18-64, and 65 and up.

The agreements for racial composition and numbers of Hispanics and Nonhispanics are, in general, well above 90 percent. This is not surprising, because there generally is a high percentage of Whites and a high percentage of Nonhispanics, and households tend to be all one race and either all Hispanic or all Nonhispanic. Because of these facts, two households of the same size picked at random will often agree in racial and Hispanic origin compositions. We still would expect that the agreement rates for racial and Hispanic origin compositions to go down as household size goes up. When different households of the same size are compared, and where there are more people, it is less likely that distributions will happen to agree. Furthermore, it is more probable that some data are missing and thus imputation necessary.

The age comparisons are interesting. There is a large difference between the frequency of agreement within 5-year age groups and for agreement within the three broader age groups. Of course, this would be true if different households were picked at random. But we believe that more is going on. It is highly improbable that two *different* households would agree in age distributions in 5-year categories. Thus, we expect that the 80 percent or so of AREX to Census households of the same size whose 5-year age distributions are the same are almost always cases where the housing units have the same people in them.

The increased agreement rate for distributions in the age groups 0-17, 18-64, and 65 and up, would represent a few cases where age was misreported by a few years, and many cases where different people were being compared, but happen to agree within these larger age groups.

Table 13. Comparisons between AREX and Census for demographic groups, for linked households with the same number of people only

HH Size	Total linked, of equal size	Equal for all sex groups [1]	Equal for all race groups	Equal for all Hisp. groups	Equal for all 5-year age groups	Equal for age groups 0-17, 18-64, 65+	Equal for all demographic groups[3]
All sizes	445,426	91.2%[2]	93.4%	94.8%	81.3%	93.1%	80.5%
1	139,292	92.2%	95.1%	97.5%	82.5%	96.1%	85.4%
2	158,259	93.8%	94.8%	95.9%	83.9%	94.0%	84.3%
3	60,641	87.1%	90.7%	92.3%	75.7%	88.4%	72.2%
4	60,181	89.3%	90.7%	90.7%	80.8%	91.7%	74.1%
5	20,723	86.8%	88.9%	89.3%	77.2%	89.0%	69.5%
6	5,359	80.4%	86.0%	86.0%	68.0%	81.8%	59.2%
7+	971	56.8%	80.8%	83.0%	28.7%	52.7%	28.7%

1. i.e., the AREX and Census households have the same number of males and the same number of females.
2. Percents are percents of the Total column.
3. Both sex groups, all race groups, both Hispanic origin groups, and age groups 0-17, 18-64, 65+.

Table 14 contains comparisons between NRFU households and other Census households of AREX and Census agreement in demographic groups. These comparisons are done by household size. In Table 10, we compare household demographic composition by household imputation status.

Table 14. Comparison of AREX and Census demographic composition of households. For linked households with the same number of people only, by size

HH Size		Total	Equal for all sex groups[1,2]	Equal for all race groups	Equal for all Hisp. groups	Equal for all 5-year age groups	Equal for age groups 0-17,18-64, 65+	Equal for all demographic groups[3]
All	NRFU	85,774	81.0%	87.7%	92.3%	58.1%	84.9%	63.4%
	non-NRFU	359,652	93.7%	94.7%	95.3%	86.9%	95.0%	84.6%
1	NRFU	31,313	82.5%	89.3%	95.7%	57.5%	91.1%	68.9%
	non-NRFU	107,979	95.0%	96.8%	98.1%	89.7%	97.5%	90.2%
2	NRFU	24,499	83.7%	88.5%	92.7%	58.6%	83.6%	64.9%
	non-NRFU	133,760	95.7%	96.0%	96.5%	88.6%	95.9%	87.9%
3	NRFU	12,549	75.7%	85.6%	89.4%	54.3%	77.1%	54.8%
	non-NRFU	48,092	90.1%	92.1%	93.0%	81.4%	91.4%	76.8%
4	NRFU	11,423	79.8%	86.3%	88.4%	63.2%	83.3%	60.2%
	non-NRFU	48,758	91.5%	91.7%	91.2%	84.9%	93.7%	77.3%
5	NRFU	4,473	78.1%	84.9%	87.2%	60.4%	80.0%	56.8%
	non-NRFU	16,250	89.2%	90.1%	89.9%	81.8%	91.4%	73.0%
6	NRFU	1,269	71.0%	80.4%	83.0%	54.0%	73.1%	46.8%
	non-NRFU	4,090	83.4%	87.8%	86.9%	72.4%	84.6%	63.0%
7+	NRFU	248	53.6%	79.8%	81.5%	27.0%	47.6%	24.6%
	non-NRFU	723	58.0%	81.2%	83.5%	29.3%	54.5%	30.2%

(Table 12 notes — from preceding page)

1. i.e., the AREX and Census households have the same number of males and the same number of females
2. Percents are percents of Total.
3. Both sex groups, all race groups, both Hispanic origin groups, and age groups 0-17, 18-64, 65+

Table 15. Comparison of AREX and Census demographic groups within households—For linked households with the same number of people only, by size

HH Size		Total	Equal for all sex groups	Equal for all race groups	Equal for all Hisp. groups	Equal for all 5-year age groups	Equal for age groups 0-17,18-64, 65+	Equal for all demographic groups
All	Imputed	4,784	49.6%	74.9%	91.7%	7.0%	60.7%	23.0%
	Not imputed	440,642	91.7%	93.6%	94.8%	82.1%	93.4%	81.2%

We see that the there is less AREX to Census agreement for NRFU households than for other Census households, both overall and for each size. From the analysis here, we cannot determine how much of the disagreement is due to inaccuracies in AREX, and how much is due to inaccuracies in Census for NRFU cases.

We also see that there is less agreement between AREX and Census for imputed households than for non-imputed households. This is not surprising, because we would not expect imputed households to agree with the true demographic composition household by household.

- *How do comparisons of household demographic composition vary for different kinds of NRFU households?*

In Table 16 below, we consider some "difficult" NRFU cases: imputed households, and those where data given by a proxy were used.

Table 16. Comparison of household demographic composition, by type of NRFU household. Linked, occupied housing units with the same number of people only.

Type of Census housing unit	Total	Same in all demographic groups*
Non-NRFU	359,652	304,312 (84.6%)
NRFU, not imputed or proxy	70,837	47,817 (67.5%)
NRFU, proxy response	10,672	5,622 (52.7%)
NRFU, Imputed	4,265	961 (22.5%)

* Both sex groups, all race groups, both Hispanic origin groups, and age groups 0-17, 18-64, 65+

As expected, the demographics did not agree much between AREX and Census for imputed households. Since these are imputations, we do not expect good agreement on a household by household basis. The disagreements here do not necessarily reflect inaccurate AREX data. For

the proxy case, again we have a drop in how similar the demographics are. Again, we suspect that this is due to inaccuracies on the Census side – proxy responses for demographics would not be expected to be very accurate in general.

- *Summary of comparisons of linked households:*

AREX households were linked with 84.1 percent of Census occupied households. AREX had the same household size for 52.1 percent of the linked occupied households. Of those occupied, linked households with the same size, AREX agreed with Census in counts in all of the demographic composition: race, sex, Hispanic origin and age (in groups 0-17,18-64, 65+) in 80.5 percent of the cases. Thus, in 41.9 percent of linked households, AREX and Census agreed in household size and demographic composition.

The results were somewhat different for the Census NRFU universe. We saw that AREX households were linked with 76.7 percent of occupied NRFU households. AREX had the same household size among occupied linked households for 38.7 percent. AREX and Census agreed in demographic composition for 63.4 percent of linked occupied households with the same size. For 24.5 percent of linked occupied NRFU households, AREX agreed in size and demographic. These compare to occupied non-NRFU households, where AREX agreed with the Census in size and household demographic composition for 48.1 percent of Census linked occupied households.

4.1.8 What was the effect of later NRFU dates on coverage and AREX to Census comparisons?

For NRFU housing units, the dates on which data are entered for the household can differ. Because AREX compares worse for NRFU households, we might expect that for later NRFU dates, AREX and Census would differ more. However, we found that there is little correlation between NRFU data entry dates and match rates.

4.1.9 What is the effect of multi-unit housing units on match rates and comparisons of sizes and demographic properties?

In Table 9 above, we showed that AREX housing units were linked with about 68 percent of occupied Census housing units which were at multi-unit basic street addresses (BSAs), and about 90 percent of occupied Census housing units which were at single-unit BSAs. Table 17 contains data regarding comparisons of coverage rates, household size, and demographic characteristics for multi-unit BSAs compared to those at single-unit BSAs.

As noted above, the match rates for occupied Census housing units at multi-unit BSAs are at about 67 percent for all sizes, and the rates for occupied Census housing units at single-unit BSAs are at about 90 percent.

For linked households, when the Census household size is one, the household size comparison of multi-units is close to that for single units. However, for larger sizes of Census household, the AREX and Census household sizes differed more frequently for households in multi-units.

For linked households of equal size, AREX differed from Census in demographic composition more often for households in multi-units. The percentage of households which agree in demographic composition runs from about 10 to 20 less for households at multi-unit addresses than for those at single-unit addresses.

We expect that people in multi-unit addresses move more often than others. Then, due to the time lag between AREX and Census it would be more probable that the households in the housing unit are different. In that case, the sizes are more likely to be different, and when they are the same, the demographic characteristics are more likely to be different.

Table 17. Comparison of match rates and household comparisons between occupied housing units at multi-unit BSAs and housing units at single-unit BSAs

Census HH Size	Group	Total	Linked (% of Total)	Equal size (%)[1]	Equal in all demographic groups(%)[2]
All sizes[3]	In multi-unit	278,447	188,826 (67.8%)	88,517 (46.9%)	64,992 (73.4%)
	In single-unit	738,826	665,915 (90.1%)	356,909 (53.6%)	293,720 (82.3%)
1	In multi-unit	135,833	91,051 (67.0%)	57,218 (62.8%)	44,978 (78.6%)
	In single-unit	140,757	125,568 (89.2%)	82,074 (65.4%)	74,034 (90.2%)
2	In multi-unit	80,719	55,820 (69.2%)	21,788 (39.0%)	15,009 (69.3%)
	In single-unit	250,753	226,676 (90.4%)	136,471 (60.2%)	118,386 (86.7%)
3-4	In multi-unit	51,244	35,165 (68.6%)	8,567 (24.4%)	4,459 (52.0%)
	In single-unit	262,714	237,644 (90.5%)	112,255 (47.2%)	83,906 (74.7%)
5-6	In multi-unit	9390	6,063 (64.6%)	926 (15.3%)	456 (49.2%)
	In single-unit	73,253	65,838 (89.9%)	25,156 (38.2%)	17115 (68.0%)
7+	In multi-unit	1,261	727 (57.7%)	18 (2.5%)	0
	In single-unit	11,349	10,189 (89.8%)	953 (9.4%)	279 (29.3%)

[1] Percent of linked households, [2] Percent of linked households of equal size, [3] Except size zero

4.1.10 What are the effects of household demographic characteristics on the match rate and AREX to Census comparisons?

- *What is the effect of household age characteristics on coverage and on comparisons between AREX and Census?*

The discrepancies between AREX and Census are partly because some households have moved out of , and others moved into, addresses between the time of AREX data and Census. For this reason, we expect that households less likely to move will have a better AREX to Census comparison. And we expect that households containing only older people are less likely to move. In Table 18, we give match rates by whether the housing unit is at multi-unit BSA, and by whether it has only people 50 and over. In Table 16, we give comparisons of match rates, size, and demographics for housing units containing only people 50 and over, and others – controlling for household size. Tables B.16 and B.17A-B in Appendix B contain similar comparisons for ages 18 and over, and for 65 and over.

The coverage by AREX of Census households with everyone over 50 are slightly, but consistently, higher. This is true whether controlling for multi-units or controlling for size. The comparison for household size and demographics are much better for households with everyone over 50, as would be expected if fewer of these households moved. (The demographic comparison is worse for households of size 3 or more, but there are few of those.)

From tables B.16 and B.17A-B in Appendix B, we see that these patterns do not hold for households with everyone over 18 compared to others. For households with everyone over 65, we see a pattern similar to that for everyone over 50.

Table 18. Coverage by multi vs. single unit, and by household age characteristics

Type of housing unit	Census household age characteristic	Total	Percent linked
All HUs	All 50 or older	292,091	85.8%
	Some under 50	725,182	83.3%
In multi-unit	All 50 or older	81,480	69.8%
	Some under 50	196,967	67.0%
In single-unit	All 50 or older	210,661	91.8%
	Some under 50	528,215	89.4%

Table 19. AREX to Census comparisons by size of housing unit and by household age characteristics

Size of HH	Census household age characteristic	Total	Linked with AREX housing units (% of Total)	Equal size (%)[1]	Equal in all demographic groups[2] (%)[3]
1	All 50 or over	148,355	121,781	86,518	78,500
			(82.1%)	(71.0%)	(90.7%)
	Some under 50	128,235	94,838	52,774	40,512
			(74.0%)	(55.6%)	(76.8%)
2	All 50 or over	137,758	123,412	83,662	76,685
			(89.6%)	(67.8%)	(91.7%)
	Some under 50	193,714	159,084	74,597	56,800
			(82.1%)	(46.9%)	(76.1%)
3+	All 50 or over	5878	5,357	2542	2072
			(91.1%)	(47.4%)	(81.5%)
	Some under 50	403,233	350,269	145,313	104,143
			(86.9%)	(41.5%)	(71.7%)

1 Percent of linked households

2 Equal in: both sex groups, all four race groups, both Hispanic origin categories, and age groups 0-17, 18-64, 65+

3 Percent of linked of equal size

- *What is the effect of the household race and Hispanic origin characteristics on match rates and comparisons between AREX and Census?*

Table 14 shows how coverage, size comparisons, and race comparisons, vary with whether there is a someone who is not White in the household according to Census.

For Census households with at least one person who is not White, the coverage by AREX is smaller, but not smaller by much, compared with other households. The fact that these coverage rates are so similar is promising. On the other hand, the household size comparisons and the racial composition comparisons display more disagreement for households that do not contain only Whites.

If we were to use administrative records for nonresponse substitution, we would want comparisons to be more similar among households with different racial characteristics. Because our administrative records appear to be more accurate and complete for Whites than for others, we should seek administrative records that have more complete and accurate data for people of other races.

Table 20. The effect of the presence of other races in a household on household match rates and comparisons

Census HH Size	Household type	Total	Linked with AREX housing units (% of Total)	Equal size (%)[1]	Equal in all four race groups (%)[2]
All sizes	all White	739,631	631,606 (85.4%)	358,833 (56.8%)	347,592 (96.9%)
	not all White	277,642	223,135 (80.4%)	86,593 (38.8%)	68,356 (78.9%)
1	all White	205,139	165,098 (80.5%)	111,112 (67.3%)	108,450 (97.6%)
	not all White	71,451	51,121 (72.1%)	28,180 (54.7%)	24,049 (85.3%)
2	all White	256,496	221,806 (86.5%)	133,180 (60.0%)	130,033 (97.6%)
	not all White	74,976	60,690 (80.9%)	25,079 (41.3%)	19,995 (79.7%)
3-4	all White	218,894	192,772 (88.1%)	93,694 (48.6%)	89,491 (95.5%)
	not all White	95,064	80,037 (84.2%)	27,128 (33.9%)	20,105 (74.1%)
5-6	all White	53,112	46,707 (87.9%)	20,319 (43.5%)	19,144 (94.2%)
	not all White	29,531	25,194 (85.3%)	5,763 (22.9%)	3,896 (67.6%)
7+	all White	5,990	5,223 (87.2%)	528 (10.1%)	474 (89.8%)
	not all White	6,620	5,693 (86.0%)	443 (7.8%)	311 (70.2%)

[1] Percent of linked households

[2] Percent of linked households of equal size

The AREX coverage of Census does not vary much with whether the household contains Hispanics. This indicates that Hispanics do not have a strong tendency to live at kinds of addresses that are hard to match, nor that our ability to get administrative records from households with Hispanics is not greatly worse.

There is a notable difference in household size comparisons between households with Hispanics and those without. This is true even controlling for Census household size.

There is a large difference between households with Hispanics and those without in Hispanic origin comparisons. Where there is a Hispanic in the household according to Census, AREX and Census agree in household Hispanic origin composition about 50 percent of the time, as compared to 96.9 percent for households with no Hispanics.

We believe that this discrepancy is largely due to imputations on the AREX side. Hispanic origin was imputed for 96.7 percent of AREX people. An imputation model would assign non-Hispanic to a high percentage of people. Because of this, we would expect that Census households composed of only non-Hispanics would agree with AREX household Hispanic origin composition a large percentage of the time. On the other hand, when a Census household does have Hispanics, we would not expect an imputation model to agree as often on a household by household basis.

Table 21. The effect of presence of Hispanics on household match rates and comparisons

Census HH Size	Household type	Total	Linked with AREX housing units (% of Total)	Equal size (%)[1]	Equal # of Hispanics (%)[2]
All sizes	All Nonhispanic	955.253	803,272 (84.1%)	424,867 (52.9%)	411,698 (96.9%)
	At least one Hispanic	62,020	51,469 (83.0%)	20,559 (39.9%)	10,365 (50.4%)
1	All Nonhispanic	268,888	210,745 (78.4%)	136,114 (64.6%)	134,063 (98.5%)
	At least one Hispanic	7,702	5,874 (76.3%)	3,178 (54.1%)	1,802 (56.7%)
2	All Nonhispanic	314,443	268,371 (85.3%)	151,588 (56.5%)	147,697 (97.4%)
	At least one Hispanic	17,029	14,125 (82.9%)	6,671 (47.2%)	4,053 (60.8%)
3-4	All Nonhispanic	287,700	250,589 (87.1%)	112,467 (44.9%)	106,922 (95.1%)
	At least one Hispanic	26,258	22,220 (84.6%)	8,355 (37.6%)	3,609 (43.2%)
5-6	All Nonhispanic	73,483	64,212 (87.4%)	23,831 (37.1%)	22,235 (93.3%)
	At least one Hispanic	9,160	7,689 (83.9%)	2,251 (29.3%)	876 (38.9%)
7+	All Nonhispanic	10,739	9,355 (87.1%)	867 (9.3%)	781 (90.1%)
	At least one Hispanic	1,871	1,561 (83.4%)	104 (6.7%)	25 (24.0%)

[1] Percent of linked
[2] Percent of linked of equal size

4.1.11 What are the effects of AREX imputations on the similarity of AREX and Census demographic characteristics?

Race and Hispanic origin were imputed for AREX people when missing (with a few exceptions). Many of our administrative records treated Hispanic as a race. Thus, we typically have either a race or a Hispanic origin for a person but not both. There are 2,282,401 people in the AREX non-GQ population. Race was imputed for 15.2% of them.

- *What are the effects of AREX race imputations on the similarity of distributions of AREX and Census race characteristics?*

In Table 10 above, and in Tables B.7 to B.11 in Appendix B, the distribution of household race characteristics for AREX is compared to that of Census, by whether or not the household had anyone whose AREX race was imputed.

From Table 10, we see that the percent of mixed race households for all AREX households is higher than that for all Census households *and* for AREX households which contained no person with an AREX imputed race. Thus, it appears that race imputation created a higher percentage of multi-race households than were indicated by Census.

4.1.12 What are the effects of race imputation on race comparisons within linked households?

Table 22 concerns linked households in which no person's AREX race was imputed, and those in which at least one person's race was imputed. The comparison is done with regard to the racial composition of the household.

The degree of agreement when races were not imputed is promising for administrative records. The fact that 78 percent of linked households with equal size required no race imputation, together with the 96 percent agreement rate in racial composition for those households shows good potential for use of administrative records for nonresponse substitution, when households are linked and have equal sizes. The percentage of agreement in household racial composition as a percentage of all of the 889,638 linked households is 37.1 percent.

As expected, when households with imputed race are included in the analysis, there is less household by household agreement between AREX and Census about racial characteristics. We consider the 86 percent agreement when race is imputed to be quite good. Since data shown above concerning *distributions* of racial composition suggested improvements in the model used to impute race, we expect that this agreement rate could be improved.

Table 22. The effect of AREX imputed race on household comparisons

Census HH Size	Total linked, with equal size [1]	HHs with at least one person with AREX imputed race: Number (% of [1]) [2]	HHs with at least one person with AREX imputed race: Equal in all race categories (% of [2])	HHs with no person with AREX imputed race: Number (% of [1]) [3]	HHs with no person with AREX imputed race: Equal in all race categories (% of [3])
All sizes*	445,426	100,416 (22.5%)	86,290 (85.9%)	345,010 (77.5%)	329,658 (95.6%)
1	139,292	5,197 (3.7%)	4,099 (78.9%)	134,095 (96.3%)	128,400 (95.8%)
2	158,259	14,087 (8.9%)	11,351 (80.6%)	144,172 (91.1%)	138,677 (96.2%)
3-4	120,822	61,389 (50.8%)	53,689 (87.5%)	59,433 (49.2%)	55,907 (94.1%)
5-6	26,082	18,991 (72.8%)	16,558 (87.2%)	7,091 (27.2%)	6,482 (91.4%)
7+	971	752 (77.4%)	593 (78.9%)	291 (22.6%)	192 (87.7%)

* Not including zero

Because there were so few people whose Hispanic origin was not imputed, we did not include a similar analysis for the effect of Hispanic origin imputation.

4.1.13 Summary of descriptive analyses

A summary of the AREX to Census comparisons is given in Table 17 below.

- *What do the results here show about the general similarity between AREX data and Census data?*

The overall coverage of occupied Census housing units by AREX was about 84 percent. For purposes of an administrative records census, the remaining 16 percent may not be of great concern. These, along with many of the vacant housing units, would require a non response operation. Note that the Census NRFU occurs *after* all mailout/mailback operations are completed. However for an administrative records census, a mailout operation would be part of the non response operation, which would make the number of cases needing phone and/or field operations even smaller.

In addition, we expect that the match rate between Census addresses and administrative records addresses could be improved by resolving many-to-one matches, improving computer match technology, and obtaining more and better quality administrative records.

Of the occupied Census linked households, AREX and Census had the same number of people in 52.1 percent of the cases. In 41.9 percent of occupied linked households, AREX and Census had the same number of people, and the same demographic composition (using the three age

categories). Relaxing the criteria somewhat, we saw that in 79.4 percent of linked households (including Census vacants), the AREX person count was within one of the Census count.

- *What do the results here show about the potential use of administrative records for nonresponse substitution?*

The under coverage of the Census universe by AREX is of more concern for purposes of nonresponse substitution. For these purposes, administrative records only, not including nonresponse operations, would probably be used. Thus, the 84 percent coverage rate is of some concern, and the fact that coverage dropped to about 77 percent for Census NRFU housing units and about 63 percent for imputed households is of more concern.

Among linked occupied households in NRFU, AREX had the same count as Census in 38.7 percent of the cases. AREX and Census had the same demographic composition for 24.5 percent of these linked occupied households. Relaxing the criteria, we saw that the AREX count was within one of the Census count for 69.3 percent of the cases, including Census vacant housing units.

We should note that Census data for NRFU are probably worse than for other households, so for some of the AREX to Census disagreement AREX may be correct.

Table 23. Summary of match rates and household comparisons between AREX and Census

Type of Housing Unit	All of Census	NRFU	non-NRFU	Imputed HHs	non-Imputed HHs
Total Occupied Census Housing Units	1,017,273	289,224	728,049	23,811	993,462
Census Occupied, linked	854,741 (84.0%)[1]	221,909 (76.7%)	632,832 (86.9%)	15,043 (63.2%)	839,698 (84.5%)
Linked occupied with equal number	455,426 (52.1%)[2]	85,774 (38.7%)	359,652 (56.8%)	4,784 (31.8%)	440,642 (52.5%)
AREX and Census counts both sex categories	406,349 (91.2%)[3]	69,488 (81.0%)	336,861 (93.7%)	2,373 (49.6%)	403,976 (91.7%)
AREX and Census counts equal in all race categories	415,948 (93.4%)[3]	75,262 (87.7%)	340,686 (94.%)	3,583 (74.9%)	412,365 (93.6%)
AREX and Census counts equal in both Hispanic origin categories	422,063 (94.8%)[3]	79,146 (92.3%)	342,917 (95.4%)	4,388 (91.7%)	417,675 (94.8%)
AREX and Census counts equal in all 5-year age categories	362,202 (81.3%)[3]	49,833 (58.1%)	312,369 (86.9%)	335 (7.0%)	361,867 (82.1%)
Equal in age groups 0-17, 18-64, 65+	414,668 (93.1%)[3]	72,835 (84.9%)	341,833 (95.1%)	2,905 (60.7%)	411,763 (93.5%)
AREX and Census counts qual in sex, race, Hispanic origin, and 5-year age groups	333,577 (74.9%)[3]	43,210 (50.4%)	290,367 (80.7%)	138 (2.9%)	333,439 (75.7%)
AREX and Census equal in demographic composition: sex, race, Hispanic origin, and age groups 0-17, 18-64, 65+	358,712 (80.5%)[3]	54,400 (63.4%)	304,312 (84.6%)	1,099 (23.0%)	357,613 (81.2%)

1. Percent of Census occupied housing units
2. Percent of Census linked housing units
3. Percent of linked housing units with equal numbers of people

4.2 Predicting Where An AREX Household Will Be Similar To A Census Household

4.2.1 Why do we need to predict where an AREX household will be similar to a census household?

While earlier analyses show that for some households the AREX household demographics are comparable to the census household demographics, in future censuses, we will not know, *before the fact*, when an AREX household will be similar to a census household. In order to effectively substitute AREX data for Nonresponse Followup data, we must be able to accurately identify the properties of addresses where AREX data are most likely to be similar to census data. The predictive model developed here is designed with this goal in mind.

4.2.2 What kind of predictive model are we fitting?

The dependent variable for this analysis is a 0-1 variable, with 1 denoting that the AREX household matched the Census household on all demographic distributions, and a zero indicating that at least one demographic distribution of the AREX household did not match the Census household. Thus the most natural form of analysis is logistic regression; in effect we will be using right hand side predictor variables to predict the probability that the two addresses will have the same demographic distribution.

4.2.3 What information was assumed and how were variables chosen to make this prediction?

The most important assumption is that, in future censuses, we would *not* have census response data available on a particular address: The only information we have is from the AREX database itself and the Master Address File. Conceivably, for future censuses, we could have tract level data from the American Community Survey; however, while we believe that this would improve our ability to predict matching demographics, we have not (yet) incorporated any similar data into this analysis.

Because the purpose of this model is to maximize predictive accuracy, we developed several hypotheses about which kinds of addresses would be most likely to match on demographic characteristics. In particular, we hypothesize:

- Nonmoving households are more likely to be captured accurately by administrative records than moving households;
- Households filing tax returns are more likely to be captured accurately than non-tax-filing households;
- Medicare households are more likely to be captured accurately than non-Medicare households;
- Households whose characteristics are corroborated by more AREX source files will be more likely to match than households with more limited corroboration among source files;
- The characteristics of households that make them "difficult to enumerate" in the census will also tend to make them "difficult to enumerate" via administrative records; therefore, mailout/mailback responders will be more likely to be captured accurately by administrative records, followed by early Nonresponse Followup responders, and so on;
- Due to the limitations on the ability of administrative records to accurately cover children, and determine their race, households with children will be less accurately captured than households without children.

In order to maximize the descriptive information in these models, various additional factors have been extracted from the AREX database and an April extract of Geography Division's Decennial Master Address File. We will comment on these additional factors at points.

4.2.4 What simple relationships occur?

We begin by describing simple bivariate relationships between data in the AREX and Master Address File and the Match/Non-match indicator. (Recall that we use the term "Demographics

match" for the "match" definition described earlier—across the age (in five year increments), race (four races), sex (two sexes), and Hispanic origin (Hispanic and non-Hispanic) array, all the characteristics of the two households are the same)[8]. In tables 21-50 that follow, cell counts will be accompanied by column percents. The most relevant two cells to compare are those where the addresses demographically match. These two cells will be marked in gray, and, in general, a large difference between the two column percents indicates that the variable listed along the top of the table is a good variable for discriminating between Match and Non-match status.

- *General properties (Colorado, NRFU status, multi-unit status).*

Table 24. AREX address location and demographic match/non-match status

	AREX address is in:		
	MD	**CO**	**Total**
Non-match	306,141	241,203	547,344
	62.5%	60.5%	
Match	184,754	157,540	342,294
	37.6%	39.5%	
Total	490,895	398,743	889,638
	55.2%	44.8%	100

We first examine basic differences between Maryland and Colorado. As can be seen above, addresses in Colorado had a slightly higher demographic match rate (39.5 percent) than addresses in Maryland (37.6 percent).

Table 25. Address Nonresponse Followup (NRFU) status and demographic match/non-match status

	From Census 2000, address is:		
	Non-NRFU	**NRFU**	**Total**
Non-match	343,267	204,077	547,344
	54.2%	79.7%	
Match	290,349	51,945	342,294
	45.8%	20.3%	
Total	633,616	256,022	889,638
	71.2%	28.8%	100%

As can be seen, addresses that are cut for the NRFU universe are much less likely to match demographically (20.3 percent) than those that are in the non-NRFU universe (45.8 percent). This suggests that administrative records data will be less useful for NRFU substitution use than originally hoped, although it will require the multivariate analysis of the next section to answer

[8] For example, the two addresses have exactly the same number of people *and*, further, they have exactly the same number of 15-19 year old Black Hispanic males, Black Hispanic Females, etc. etc.

the question: Is the lower demographic match rate of NRFU addresses a result of their characteristics, or is there a fundamental problem with NRFU addresses?

Table 26. Single unit or multi unit address (from Census 2000 HDF) and demographic match/non-match status

	From Census 2000 data: Single unit at BSA	Multi unit at BSA	Total
Non-match	406,986	140,358	547,344
	59.1%	69.8%	
Match	281,486	60,808	342,294
	40.9%	30.2%	
Total	688,472	201,166	889,638
	77.4%	22.6%	100%

As can be seen, addresses that are identified as multiple addresses at a BSA by the MAF are less likely to match demographically (30.2 percent) versus those that are single unit (only one address at a BSA [40.9 percent]).

Table 27. Single unit or multi unit address (from AREX) and demographic match/non-match status

	From AREX data: Single unit at BSA	Multi unit at BSA	Total
Non-match	413,638	133,706	547,344
	59.3%	69.5%	
Match	283,566	58,728	342,294
	40.7%	30.5%	
Total	697,204	192,434	889,638
	78.4%	21.6%	100%

A similar effect occurs for addresses that are identified as multiple addresses at a BSA by administrative records data. Addresses with multiple units match 30.5 percent of the time; addresses with single units match 40.7 percent of the time. We note for the record, however, that this result could occur because of difficulties caused by Census operations, for example, misdeliveries of census forms to incorrect apartments, rather than administrative records.

Table 28. Number of units at BSA (from AREX) and demographic match/non-match status

	From AREX data: Less than 10 units at BSA	From AREX data: 10 or more units at BSA	Total
Non-match	466,198	81,146	547,344
	60.7%	66.5%	
Match	301,487	40,807	342,294
	39.3%	33.5%	
Total	767,685	121,953	889,638
	86.3%	13.7%	100%

For addresses with ten or more units at the BSA, we see that these addresses are less likely to match demographically (33.5 percent) than those that have one to nine units at the BSA (39.3 percent). However, we note that 33.4 percent is actually slightly higher than the previous table (30.5 percent)—addresses with ten or more units are the BSA are slightly more likely to match demographically than addresses that are multiunit in general.

In the next section, we explore whether characteristics of the administrative records address can explain demographic matching. For example; are addresses that come from particular source files more likely to match demographically than records that do not?

4.2.5 Source Files.

In this section, we will determine that the source file of an address bears a relationship with its match/non-match status. When we refer to an address as being "in" a file (for example, an address "in the IRS 1040 file"), we mean the following: At least one person determined to reside at that address by AREX processing had their address come from the specified file. A single address could be "in" multiple source files by virtue of the persons determined to reside at that address coming from different source files.

Table 29. Address is found in the IRS 1040 file versus demographic match/non-match status

	Not in IRS	In IRS file	Total
Non-match	133,291	414,053	547,344
	73.6%	58.5%	
Match	47,932	294,362	342,294
	26.4%	41.6%	
Total	181,223	708,415	889,638
	20.4%	79.6%	100%

As can be seen, for addresses in which at least one person at that address came from the IRS 1040 file, 41.6 percent of these addresses matched demographically, as opposed to 26.4 percent where no person at the address was found on the IRS 1040 file. This suggests that presence on the IRS file is a predictor of accurate demographic matching.

Table 30. Address is found in the HUD-TRACS file versus demographic match/non-match status

	Not in HUD	In HUD file	Total
Non-match	547,308	36	547,344
	61.5%	100%	
Match	342,294	-	342,294
	38.5%	0%	
Total	889,602	36	889,638
	100%	0%	100%

As can be seen, so few addresses came only from the HUD TRACS file, that no substantive inferences can be made, except perhaps to note that none of them demographically matched their census counterparts.

Table 31. Address is found in Medicare versus demographic match/non-match status

	Not in Medicare	In Medicare	Total
Non-match	456,058	91,286	547,344
	66.1%	45.7%	
Match	233,619	108,675	342,294
	33.9%	54.3%	
Total	689,677	199,961	889,638
	77.5%	22.5%	100%

As can be seen, addresses for which at least one person came from the Medicare file matched at a notably higher rate (54.3 percent) than addresses in which no one came from the Medicare file (33.9 percent). The difference between these two percentages (about 21 percent) is one of the largest that we will find, suggesting that presence on the Medicare file is a substantial predictor of demographic matching. Our later multivariate analyses will question this relationship somewhat, however.

Table 32. Address is found in Information Returns Master File (IRMF) versus demographic match/non-match status

	Not in IRMF	In IRMF	Total
Non-match	117,382	429,962	547,344
	81.2%	57.7%	
Match	27,210	315,084	342,294
	18.8%	42.3%	
Total	144,592	745,046	889,638
	16.3%	83.7%	100%

As can be seen, addresses for which at least one person came from the Information Returns Master File (IRMF) matched at a higher rate (42.3 percent) than those for which no one come from the IRMF (18.8 percent). The difference between these two percentages(about 24 percent)

is again one of the largest that we will find, suggesting that the non-presence on the IRMF is a substantial predictor of an address not demographically matching.

Table 33. Address is found in Indian Health Service (IHS) versus demographic match/non-match status

	Not in IHS file	In IHS file	Total
Non-match	547,201	143	547,344
	61.5%	83.6%	
Match	342,266	28	342,294
	38.5%	16.4%	
Total	889,467	171	889,638
	100.0%	0.0%	100%

Only a small number of addresses came from the IHS file in the AREX test sites; and, for those that did, they demographically matched at a lower rate (16.4 percent) than those that did not (33.5 percent). Thus, presence on this file is a predictor of the addresses demographically ***not*** matching.

Table 34. Address is found in the Selective Service System (SSS) versus demographic match/non-match status

	Not in SSS	In SSS	Total
Non-match	508,423	38,921	547,344
	60.4%	82.3%	
Match	333,898	8,396	342,294
	39.6%	17.7%	
Total	842,321	47,317	889,638
	94.7%	5.3%	100%

47,317 addresses had one or more persons coming from the Selective Service file. However, those that did had a lower rate of demographic matching (17.7 percent) than those that did not (39.6 percent).

The results of the IRS, IRMF, and Medicare tables above led us to explore the following two way interactions between IRS and IRMF, IRMF and Medicare, and Medicare and IRS.

Table 35. Address is found in both IRS 1040 and IRMF versus demographic match/non-match status

	Not in IRS& IRMF	In IRS & IRMF	Total
Non-match	180,355	366,989	547,344
	73.9%	56.8%	
Match	63,728	278,566	342,294
	26.1%	43.2%	
Total	244,083	645,555	889,638
	27.4%	72.6%	100%

As can be seen, addresses that are in both the IRS 1040 and the IRMF are more likely to match demographically (43.2 percent) than addresses that are in neither (26.1 percent). This is slightly higher than either individually, but only slightly.

Table 36. Address is found in both IRS 1040 and Medicare versus demographic match/non-match status.

	Not in IRS & Medicare	In IRS & Medicare	Total
Non-match	478,492	68,852	547,344
	64.6%	46.4%	
Match	262,754	79,540	342,294
	35.4%	53.6%	
Total	741,246	148,392	889,638
	83.3%	16.7%	100%

As can be seen, addresses that are in both the IRS 1040 and the Medicare are more likely to match demographically (53.6 percent) than addresses that are in neither (35.4 percent). This is slightly lower than the Medicare only table above (54.3 percent), suggesting that, conditional on knowing that an address came from the Medicare file, knowing that it also came from the IRS 1040 does not provide any additional predictability about its demographic match.

Table 37. Address is found in both IRMF and Medicare versus demographic match/non-match status

	Not in IRMF & Medicare	In IRMF & Medicare	Total
Non-match	459,628	87,716	547,344
	66.0%	45.4%	
Match	236,650	105,644	342,294
	34.0%	54.6%	
Total	696,278	193,360	889,638
	78.3%	21.7%	100%

As can be seen, addresses that are in both the IRMF and Medicare are more likely to match demographically (54.6 percent) than addresses that are in neither (34.0 percent). This is slightly

higher than the Medicare only table above (54.3 percent), but only very slightly. Again this suggests that, conditional on knowing that an address came from the Medicare file, knowing that it also came from the IRMF does not provide any additional predictability about its demographic match.

Finally, we present the results of being in all three files.

Table 38. Address is found in IRS 1040, IRMF, and Medicare versus demographic match/non-match status

	Not in IRS, IRMF & Medicare	In IRS, IRMF and Medicare	Total
Non-match	479,450	67,894	547,344
	64.5%	46.2%	
Match	263,387	78,907	342,294
	35.5%	53.8%	
Total	742,837	146,801	889,638
	83.5%	16.5%	100%

As can be seen, addresses that are in all three files are more likely to match demographically (53.8 percent) than addresses that are in none (35.5 percent). This is slightly lower than the Medicare only table above (54.3 percent). Again this suggests that, conditional on knowing that an address came from the Medicare file, knowing that it also came from the IRMF and IRS 1040 does not provide any additional predictability about its demographic match.

The next section focuses on variables that we hypothesized, and later exploratory analysis confirmed, that predict demographic matching.

4.2.6 Demographic properties (Size, age, race, imputation status).

Table 39. Number of persons in the AREX address versus demographic match/non-match status

	AREX household number of persons							
	0	1	2	3	4	5	6+	Total
Not Matched	66,939	106,529	119,419	105,101	78,193	39,826	31,337	547,344
	88.1%	49.3%	48.9%	72.2%	65.2%	74.8%	91.0%	
Matched	9,011	109,680	124,895	40,475	41,756	13,390	3,087	342,294
	11.9%	50.7%	51.1%	27.8%	34.8%	25.2%	9.0%	
Total	75,950	216,209	244,314	145,576	119,949	53,216	34,424	889,638

The effect of number of people on the AREX file is distinctive: Essentially, those addresses with exactly one or two persons in the administrative records database are much more likely to match demographically than those addresses that have zero, three, or more persons. This suggests that administrative records data will tend to match demographically more with smaller households than with larger. This is confirmed when we collapse the above table.

Table 40. One or Two persons in AREX address versus demographic match/non-match status

	AREX household has only one or two persons		
	No	**Yes**	**Total**
Non-match	321,396	225,948	547,344
	74.9%	49.1%	
Match	107,719	234,575	342,294
	25.1%	51.0%	
Total	429,115	460,523	889,638
	48.2%	51.8%	100%

Confirming the previous table, addresses with only one or two persons on the administrative records file are more likely to match demographically (50.9 percent) than addresses with zero or three or more persons (25.1 percent).

Table 41. AREX imputed race versus demographic match/non-match status

	At least one AREX person has imputed race:		
	No	**Yes**	**Total**
Non-match	395,818	151,526	547,344
	58.7%	70.4%	
Match	278,689	63,605	342,294
	41.3%	29.6%	
Total	674,507	215,131	889,638
	75.8%	24.2%	100%

As can be seen, if at least one AREX person had their race imputed using the AREX 2000 race imputation rules, then that household is less likely to demographically match (29.6 percent) than addresses where no person had their race imputed (41.3 percent). This contributes further evidence (beyond that found in other AREX reports and evaluations) that the race imputation model does not work at these small levels of geography, even though it generates correct aggregate distributions.

Table 42. Address has children versus demographic match/non-match status

	AREX Children in Household?		
	No	**Yes**	**Total**
Non-match	351,560	195,784	547,344
	57.8%	69.6%	
Match	256,894	85,400	342,294
	42.2%	30.4%	
Total	608,454	281,184	889,638
	68.4%	31.6%	100%

As can be seen, if the address has at least one child, then that household is less likely to demographically match (30.4 percent) than an address where no children are believed to be present (42.2 percent). This may reflect difficulties in modeling of race for children or difficulties of accurately capturing children in administrative records.

Table 43. Address contains only persons 65 and older versus demographic match/non-match status

	All AREX persons age 65 or older?		
	No	**Yes**	**Total**
Non-match	513,926	33,418	547,344
	66.6%	28.4%	
Match	258,150	84,144	342,294
	33.4%	71.6%	
Total	772,076	117,562	889,638
	86.8%	13.2%	100%

As can be seen, having all persons in the household aged 65 or older is a very strong predictor of demographic matching (71.6 percent), as opposed to other households (33.4 percent). We do not know if this is because of better data quality for persons 65 or older, the Medicare source file for many of these persons, lower mobility rates of such addresses, or more accurate census responses by such persons. These are all conceivable explanations for this effect.

Table 44. Address contains only persons 50 and older versus demographic match/non-match status

	All AREX persons age 50 or older?		
	No	**Yes**	**Total**
Non-match	475,248	72,096	547,344
	71.3%	32.4%	
Match	191,618	150,676	342,294
	28.7%	67.6%	
Total	666,866	222,772	889,638
	75.0%	25.0%	100%

Having seen the effect in the previous table, we also wished to explore whether addresses where everyone was age 50 and older would have similar characteristics. As can be seen, the effect is somewhat less strong: Addresses where every person is 50 or older match 67.6 percent of the time, while addresses where this is not the case match 28.7 percent of the time.

The next section explores whether demographic characteristics of the address itself (taken from the administrative records files) predict demographic matching.

Table 45. AREX contains at least one White person versus demographic match/non-match status

	AREX household has at least one White person?		
	No	**Yes**	**Total**
Non-match	197,578	349,766	547,344
	78.8%	54.7%	
Match	53,217	289,077	342,294
	21.2%	45.3%	
Total	250,795	638,843	889,638
	28.2%	71.8%	100%

As can be seen, addresses with at least one White person match demographically at a higher rate (45.3 percent) than addresses that do not have at least one White person (21.2 percent).

Table 46. AREX contains at least one black person versus demographic match/non-match status

	AREX household has at least one black person?		
	No	**Yes**	**Total**
Non-match	408,668	138,676	547,344
	57.7%	76.7%	
Match	300,085	42,209	342,294
	42.3%	23.3%	
Total	708,753	180,885	889,638
	79.7%	20.3%	100%

As can be seen, addresses with at least one Black person match demographically at a lower rate (23.3 percent) than addresses that do not have at least one Black person (42.3 percent).

Table 47. AREX contains at least one American Indian person versus demographic match/non-match status

	AREX household has at least one American Indian person?		
	No	**Yes**	**Total**
Non-match	541,875	5,469	547,344
	61.3%	95.9%	
Match	342,063	231	342,294
	38.7%	4.1%	
Total	883,938	5,700	889,638
	99.4%	0.6%	100%

As can be seen, within the AREX test sites, addresses with at least one American Indian person match demographically at a substantially lower rate (4.1 percent) than addresses that do not have at least one American Indian person (38.7 percent).

Table 48. AREX contains at least one Asian or Pacific Islander person versus demographic match/non-match status

	AREX household has at least one Asian/PI person?		
	No	**Yes**	**Total**
Non-match	524,711	22,633	547,344
	60.9%	81.6%	
Match	337,177	5,117	342,294
	39.1%	18.4%	
Total	861,888	27,750	889,638
	96.9%	3.1%	100%

As can be seen, addresses with at least one Asian or Pacific Islander person match demographically at a substantially lower rate (18.4 percent) than addresses that do not have at least one Asian or Pacific Islander person (38.7 percent).

Table 49. AREX contains at least one Hispanic person versus demographic match/non-match status

	AREX household has at least one Hispanic person?			
	No	**All missing**	**Yes**	**Total**
Non-match	492,491	170	54,683	547,344
	59.5%	100%	88.5%	
Match	335,160	-	71,343	42,294
	40.5%	0%	11.5%	
Total	827,651	170	61,817	889,638
	93.0%	0.0%	6.9%	100%

In this table, we see an additional column: All missing. When developing or imputing Hispanic origin status, there existed persons where the AREX processing had almost literally no information on which to make a flag, either Hispanic or non-Hispanic. These individual person records were flagged with missing Hispanic origin. For an address full of such persons, it also receives a special "all missing" code. These represent only 170 out of the 889,638 addresses in the two test sites.

As can be seen, addresses with at least one Hispanic person match demographically at a substantially lower rate (11.5 percent) than addresses that do not have at least one Hispanic person (40.5 percent).

Table 50. All persons in the same household have the same Hispanic origin versus demographic match/non-match status

	AREX household persons all have the same Hispanic origin?			
	No	**All missing**	**Yes**	**Total**
Non-match	39,907	181	507,256	547344
	91.1%	100%	60.0%	
Match	3,877	-	338,417	342294
	8.9%	0.0%	40.0%	
Total	43,784	181	845,673	889638
	4.9%	0.0%	95.1%	100%

As can be seen, addresses with all persons of the same Hispanic origin (all Hispanic or all non-Hispanic) match demographically at a substantially higher rate (40.0 percent) than addresses that do not all have the same Hispanic origin (8.9 percent).

Table 51. All persons in the same household have the same race versus demographic match/non-match status

	AREX household persons all have the same race?			
	No	**All missing**	**Yes**	**Total**
Non-match	35,678	5,011	506,655	547344
	91.4%	100.0%	59.9%	
Match	3,348	-	338,946	342294
	8.6%	0.0%	40.1%	
Total	39,026	5,011	845,601	889638
	4.4%	0.6%	95.1%	100%

As can be seen, addresses with all persons of the same race (e.g. all Black or all Asian/Pacific Islander) match demographically at a substantially higher rate (40.1 percent) than addresses that do not all have the same race (8.6 percent).

Table 52. Hispanic origin imputation status versus demographic match/non-match status

	AREX household has at least one imputed Hispanic person?		
	No	**Yes**	**Total**
Non-match	74,523	472,821	547,344
	84.0%	59.0%	
Match	14,187	328,107	342,294
	16.0%	41.0%	
Total	88,710	800,928	889,638
	10.0%	90.0%	100%

As can be seen, addresses in which at least one person has imputed Hispanic origin match demographically at a substantially higher rate (41.0 percent) than addresses that do not all have at least one person with imputed Hispanic origin (16.0 percent). This runs directly counter to the race imputation questions, where the effect of having a person's race imputed was to reduce the matching rate.

Table 53. No AREX person has imputed race versus demographic match/non-match status

	AREX household has no one with imputed race		
	No	**Yes**	**Total**
Non-match	151,526	395,818	547,344
	70.4%	58.7%	
Match	63,605	278,689	342,294
	29.6%	41.3%	
Total	215,131	674,507	889,638
	24.2%	75.8%	100%

Addresses in which no person had imputed race are more likely to match demographically (41.3 percent) than those that had one or more persons with imputed race (29.6 percent). Clearly, race imputation was associated with non-matching on demographic characteristics.

Finally, an interesting phenomenon occurs for addresses that were found in the 1998 and 1999 LUCA (Local Update of Census Addresses) programs, which we will comment on but not elaborate here. Essentially, any address that was added during LUCA, or LUCA appeals, and was verified to exist in field verification, tended to be more likely to match demographically.

4.2.7 What multivariate relationships occur?

Based on the exploratory analysis of the previous section, we have constructed a multivariate logistic regression model. This model was not the result of a specification search; instead, variables and their coding were chosen based on their bivariate predictability, described above, and entered into a single logistic regression model (to avoid problems with multiplicity). Only variables were chosen that would be available prior to decennial Census operations, with two exceptions: an indicator that the Census address was a census Enumerator return and an indicator that the Census address was imputed. These two indicators were included to provide additional information about relative effect sizes of AREX data versus NRFU and imputation status. Nonetheless, we remind the reader that these matched households are not a representative sample from some population of households, thus, in any case, standard error estimates, z-statistics, and p-values should be considered illustrative only, and guides to future inferential modeling.

Table 54. Overall Response Profile for the "Match" Variable

Response Profile and Overall Model Fit Statistics	
Match Status	Total Frequency
Demographics Match	342294 (38.5%)
Demographics Do Not Match	547344 (61.5%)

As can be seen, 38.5 percent of all addresses that were linked during computer matching, also match on demographics. Conversely, 61.5 percent do not.

Table 55. Goodness-of-Fit Measures for the Logistic Regression Model

Criterion	Intercept Only	Intercept and Covariates
AIC	1,185,613.2	1,001,550.2
SC	1,185,624.9	1,001,831.0
-2 Log L	1,185,611.2	1,001,502.2
Pseudo R-Square	0.1869	

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	184,108.945	23	<.0001

(full model versus null model of intercept only)

Note: N=889,638 households in two AREX test sites in Colorado and Maryland whose addresses were computer linked; A household is declared "matched" if it's age, race, sex and Hispanic origin composition is the same across the AREX household and the equivalent census household. AIC is the Akaike Information Criterion; SC is the Schwarz criterion. –2 Log L is –2 times the log likelihood (LL) of the model, evaluated at its maximum; R-square is the pseudo R-square value, consisting of (LL(model) – LL(intercept only))/LL(model). The Likelihood Ratio test tests the null hypothesis that all coefficients except the intercept are zero in the population; Pr>ChiSq is the (nominal) probability of obtaining that Chi-Square value by chance; Because observations are not drawn from a probability sample from any particular population, all standard errors, Chi-square tests, and significance testing should be considered illustrative only. (Note is also applicable to Table 53).

As can be seen in all of these tests, the full model dramatically improves upon the null (intercept only) model. The Pseudo R-Square value indicates that the model results in a 19 percent improvement in the log-likelihood over the null model of an intercept only.

The following table provides maximum likelihood estimates of the full model with all interaction terms included.

Table 56. Maximum Likelihood Parameter Estimates, Standard Errors, and Approximate Tests

Row Number	Variable	df	Estimate	Standard Error	Wald Chi-Square	PR >ChiSq	Exp (Est)
[0]	Intercept	1	-2.756	0.050	2977.43	<.0001	0.064
[1]	Colorado Effect	1	-0.102	0.005	379.62	<.0001	0.903
[2]	Enumerator Return	1	-1.096	0.006	26648.72	<.0001	0.334
[3]	Imputed Return	1	-3.133	0.110	809.52	<.0001	0.044
[4]	Not Multi-unit	1	0.926	0.018	2656.05	<.0001	2.525
[5]	One or Two Persons	1	0.982	0.011	7013.33	<.0001	2.672
[6]	No Imputed Race	1	0.790	0.018	1778.60	<.0001	2.205
[7]	Hhold has Children	1	0.275	0.007	1239.27	<.0001	1.317
[8]	Hhold has 1+White	1	0.598	0.009	4168.03	<.0001	1.819
[9]	Hhold all age 65+	1	0.281	0.187	2.25	0.1334	1.325
[10]	In IRS File	1	-0.048	0.047	1.04	<0.3075	0.953
[11]	In IRMF File	1	-0.341	0.047	52.61	<.0001	0.710
[12]	In Medicare File	1	-0.076	0.048	2.50	<0.1136	0.927
[13]	In IRS & IRMF	1	0.901	0.047	363.32	<.0001	2.462
[14]	In IRS & Medicare	1	-0.488	0.015	996.77	<.0001	0.614
[15]	In Medicare and IRMF	1	0.390	0.047	68.23	<.0001	1.478
[16]	Age 65+ & One/Two	1	0.870	0.156	30.81	0.0001	2.389
[17]	Age 65+ & 1 + White	1	-1.042	0.167	38.63	<.0001	0.353
[18]	One/Two & 1 + White	1	-0.036	0.013	8.001	<0.0047	1.037
[19]	65+ & 1 or 2 & 1+ White	1	0.974	0.168	33.25	<.0001	2.649
[20]	65+ and not Multi-unit	1	-1.021	0.119	73.41	<.0001	0.360
[21]	65+ and no Imputed Race	1	0.425	0.105	16.23	<.0001	1.531
[22]	No Imp.Race and not Multi	1	-0.630	0.019	1057.22	<.0001	0.532
[23]	65+ & no Imp. Race & not Multi	1	0.657	0.120	29.90	<.0001	1.931
[10]*[11]* [13]	Total Effect of Capture in IRS and IRMF						1.666
[10]*… …*[15]	Total Effect of Capture in all Three Files						1.401
[5]*[8]* [9]*[16]… *[19]	Total Effect of all of 65+, White, and 1/2 Person Hhold						14.92
[4]*[6]* [9]*[20] …*[23]	Total Effect of all of 65+, Nonmulti-unit, nonimputed race						4.177

Note: N=889,638 households in two AREX test sites in Colorado and Maryland whose addresses were computer linked; A household is declared "matched" if its age, race, sex and Hispanic origin composition is the same across the AREX household.

This table indicates individual coefficients estimated via maximum likelihood. The rightmost column indicates exponentiated coefficients, and can be interpreted as the change in the odds of being a match given a one unit change in the independent variable, holding all other variables

constant. An exponentiated coefficient of one indicates no effect, greater than one indicates positive effect, and less than one indicates negative effect.

We will begin with individual effects (rows [1]—[9]). Households in the Colorado test site are slightly less likely to match census demographics, all other effects held constant, as indicated by the exponentiated coefficient less than one. Households where an enumerator enumerated the household (as opposed to a mailout/mailback household) are substantially less likely to match census demographics, and households where the census return was imputed are, not surprisingly, very unlikely to have the same demographics as their AREX counterparts.

Addresses that are not multiunit (that is, only a single address resides at the basic street address) are 2.53 times more likely to match census demographics, holding other effects constant. Addresses that consist of only one or two persons are 2.67 times more likely to match census demographics, holding other effects constant. Addresses that have no person with imputed race are 2.21 times more likely to match census demographics, holding other effects constant. A household that has children is slightly more likely to match census demographics, holding all other effects constant (obviously, this interacts with other variables in the model—an address with children logically cannot have all persons 65 or older). A household that has at least one person of White race is 1.82 times more likely to match census demographics, holding all other effects constant. Finally, a household will all persons 65 or older is 1.33 times more likely to match census demographics, holding other effects constant.

Because there are several two- and three-way interaction terms in the model, the remainder of the coefficients deserve special care in their interpretation. Rather than describe individual two-way and three-way interactions, we will focus on the variables' "total effect". The last four rows of the table indicate the "total effect" of combinations of variables, calculated by multiplying their exponentiated coefficients. As can be seen in the row labeled "total effect of capture in IRS and IRMF", a household with at least one person captured by IRS 1040 *and* at least one person captured by IRMF is 1.666 times more likely to match demographics than a household not so composed. The total effect of being captured in IRS, IRMF, and Medicare is 1.401 times more likely to match demographics than a household not so composed.

The total effect of having all persons 65 or older, at least one White person, and consisting only of a one or two person household is dramatically positive. A household composed of each of the above is about ***fifteen times*** more likely to match census demographics, holding other effects constant. Similarly, a household having all persons 65 or older, not being a multiunit address, and having no imputation from the administrative records is about ***four*** times more likely to match census demographics, holding other effects constant.

Table 57. Classification Results for Predicted Probabilities .5,…,.8

Classification Table									
	Correct		Incorrect		Percentages				
Prob. Level	Event	Non-Event	Non-Event	Event	Correct	Sensitivity	Specificity	False POS	False NEG
0.5	184,230	457,943	89,401	158,064	72.2	53.8	83.7	32.7	25.7
0.6	110,701	506,699	40,645	231,593	69.4	32.3	92.6	26.9	31.4
0.7	72,335	530,307	17,037	269,959	67.7	21.1	96.9	19.1	33.7
0.8	32,373	540,798	6,546	309,921	64.4	9.5	98.8	16.8	36.4

As can be seen, if we choose the cutoff of .5 (so that we predict a "match" when P[match=1|XB] is greater than or equal to .5), we obtain about 184,000 correct match predictions, and about 458,000 correct *non-match* predictions. Similarly, we obtain about 89,000 incorrect match predictions, and about 158,000 incorrect non-match predictions. This totals 72.2 percent correct predictions, 53.8 percent of the matches correctly predicted to be matches (sensitivity), 83.7 percent of the non-matches correctly predicted to be non-matches (specificity), a 32.7 percent false positive rate and a 25.7 percent false negative rate.

We need not choose .5 as our cutoff, however. If we choose a more stringent cutoff, for example .8 (so that we predict a "match" only when P[match=1|XB] is greater than or equal to .8), we obtain about 32,000 correct match predictions, about 541,000 correct non-match predictions, only 6,546 incorrect non-match predictions, and about 310,000 incorrect non-match predictions. This generates 64.4 percent overall correct predictions, but a false positive rate of only 16.8 percent, with a correspondingly higher false negative rate of 36.4 percent. Of course, by using such a stringent cutoff, we in fact *miss* most of the actual matches (sensitivity drops to 9.5 percent), but we are quite sure to correctly predict most of the actual non-matches (specificity climbs to 98.8 percent).

In order to evaluate cutoffs and their implications for goodness of fit, sensitivity and specificity, we present the following evaluative figures. Figure 2 provides an assessment of the goodness of fit of the obtained logit function against "jittered" outcomes.

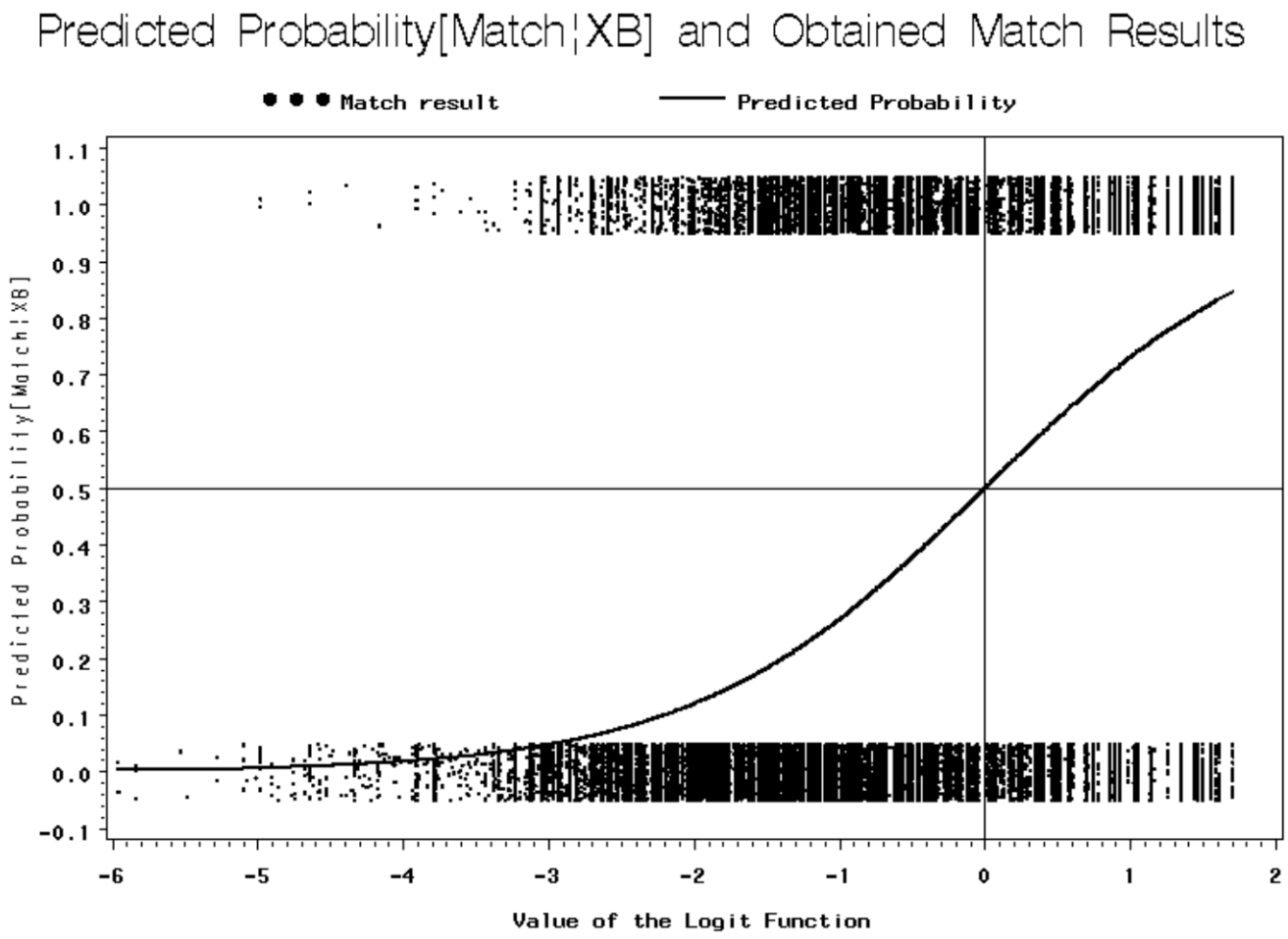


Figure 2. Goodness of Fit Diagnostic Plot

In this figure, the ordinate is the value of the logit function ln(p/1-p). A 10 percent sample of the 889,638 observations are plotted here. Each individual observation (a linked pair of addresses) is plotted as a point near zero or one. The points have been "jittered" slightly to simulate density and avoid overplotting. The abscissa is the predicted probability that an observation will be a match. If we choose .5 as our cutoff (so that we declare an observation a predicted match if P[match=1|XB]>.5), then this corresponds to a logit value of zero, and the vertical line. The horizontal line at .5 is for reference. Points in the upper right hand quadrant are "hits"—correct predictions that the demographics of the households match. Points in the lower left hand quadrant are also "hits"—correct predictions that the demographics of the households will *not* match. Points in the upper left hand and lower right hand quadrants are misses—incorrect predictions. Goodness of fit is assessed by comparing the predicted logit function to the density of the obtained match outcomes. (For more on the development and interpretation of this graph, see Judson, 1992).

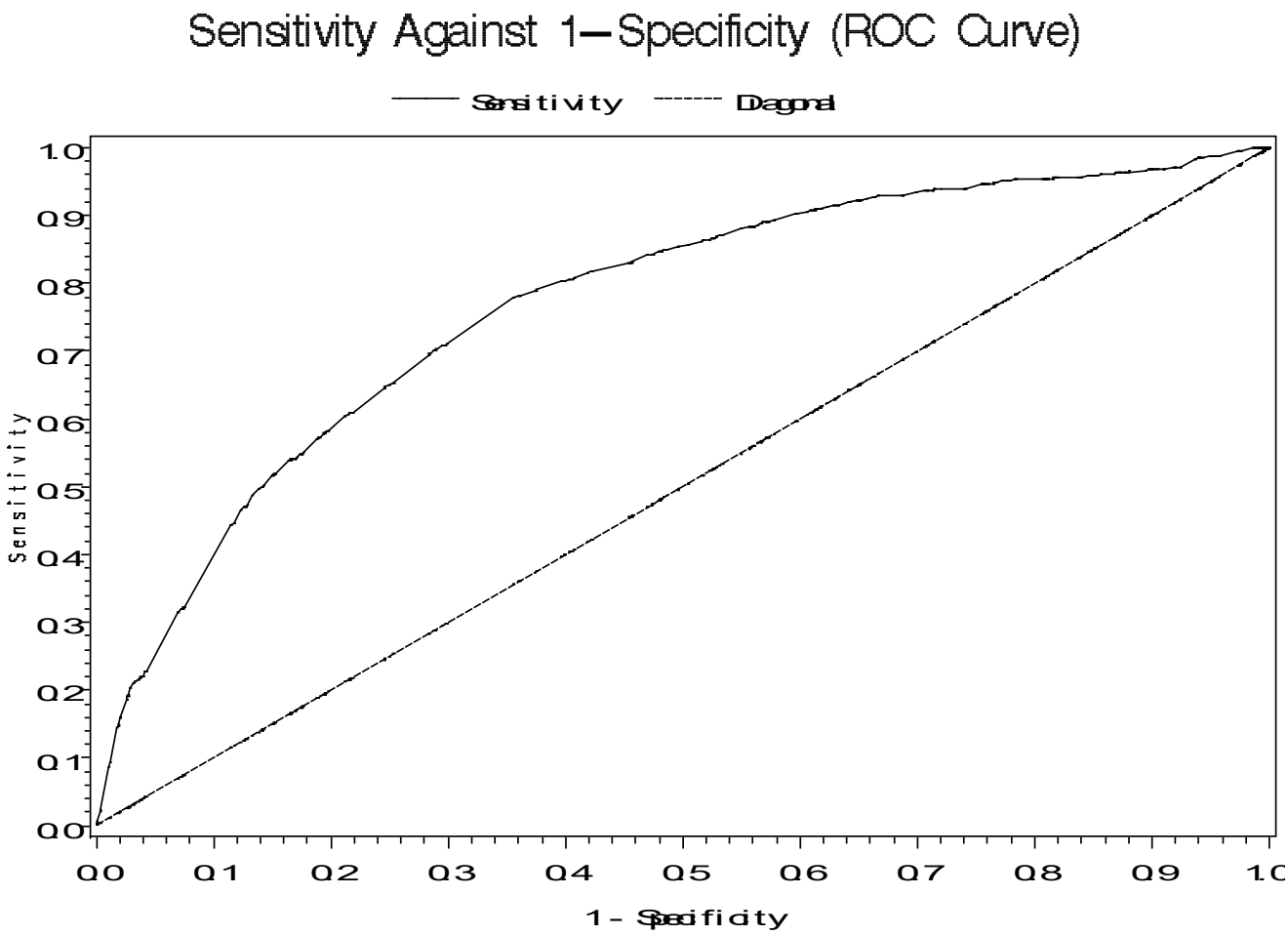


Figure 3. Receiver Operating Characteristic (ROC) curve

Figure three is a Receiver Operating Characteristic (ROC) curve, first developed in signal detection theory (Peterson, Birdsall, and Fox, 1954; Green and Swets, 1974; StataCorp, 2001). ROC curves are typically used when the point of the analysis is correct classification, as it is here. The user must specify a "cutoff" above which to declare an observation a match. The curve starts at (0,0), where the cutoff is c=1, and continues to (1,1), where the cutoff is c=0. A model with no predictive power would be at the diagonal, where sensitivity = 1-specificity, so both match and non-match cases are being predicted equally well (or poorly). The greater the predictive power of the model, the more bowed the curve. As can be seen, the curve is substantially better than the null diagonal model; however, it has some way to go to be fully bowed in the upper left hand quadrant, thus suggesting that further improvement is in order.

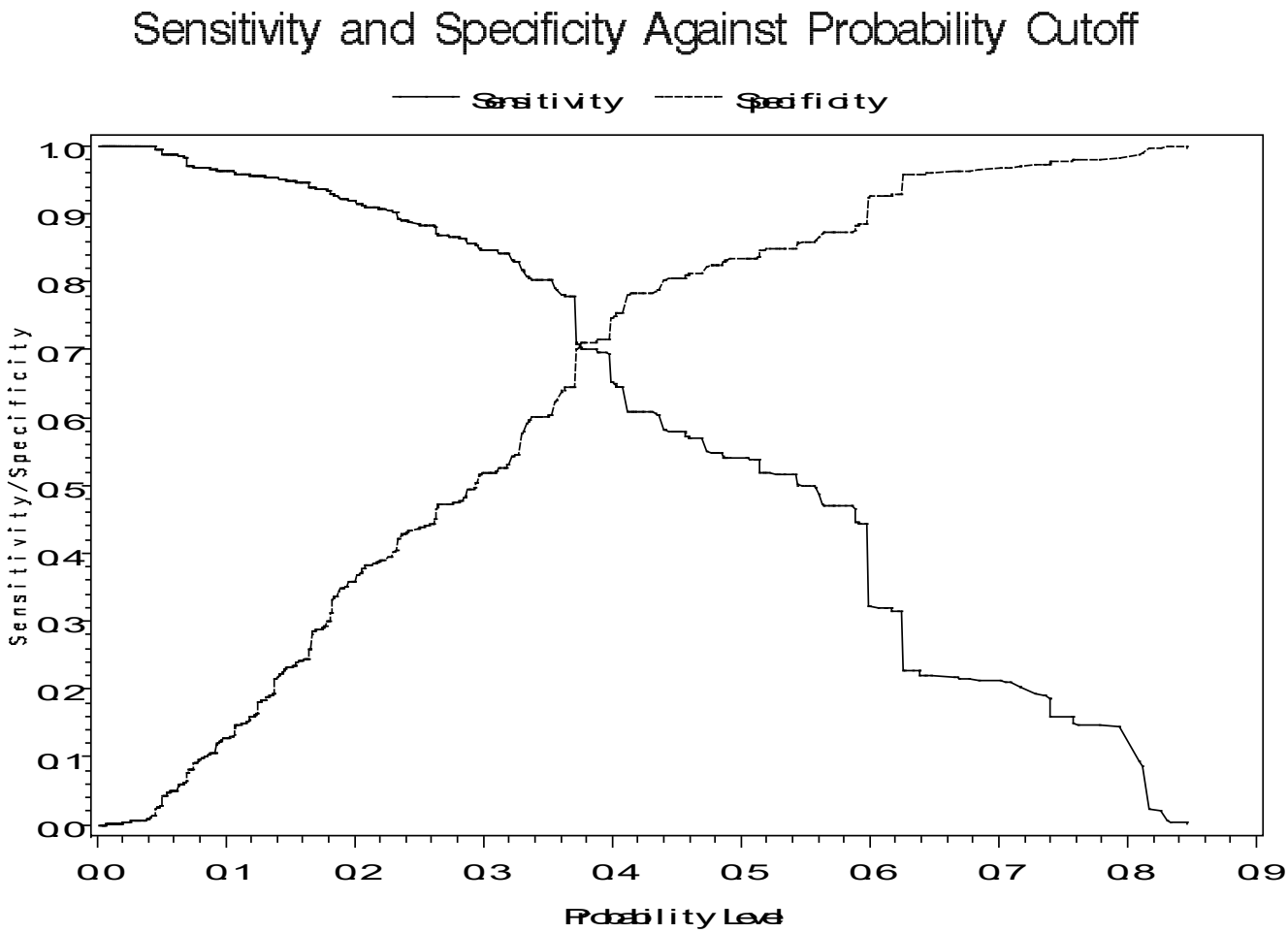


Figure 4. Plot of Sensitivity and Specificity Versus User-Chosen Probability Cutoff

Figure four plots the sensitivity and specificity directly against the user-chosen cutoff. These two curves provide a "guide" to the user as to which probability level to choose—that is, if the user chooses as cutoff value P[match|XB]=c, what sensitivity and specificities will he/she endure? An example is the cutoff value of .7: Should we require that the model predict that there is a 70 percent chance that an observation has matched demographics before we so make that prediction, then we will successfully detect about 21 percent of the true matches, and successfully detect about 95 percent of the true non-matches. If we choose a cutoff of .5 for this decision, we will successfully detect about 50 percent of the true matches, but only 80 percent of the true non-matches will be successfully detected. Obviously, we want high sensitivity and high specificity, but we cannot get both.

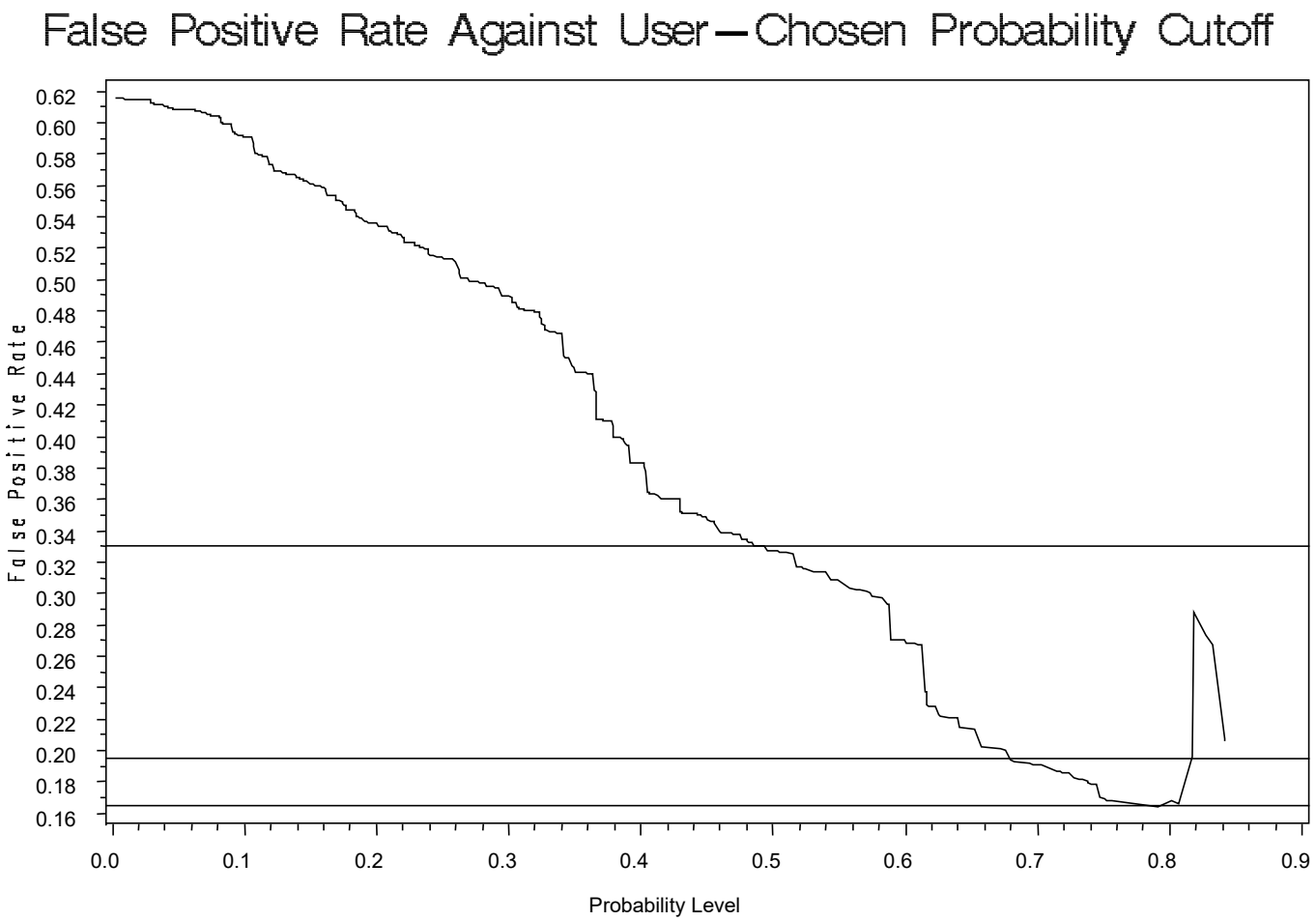


Figure 5. False Positive Rate Versus User-Chosen Probability Cutoff

Finally, a relevant diagnostic for the classification problem is to plot the false positive rate (= # false positives/ (# false positives + # correct positives) for a given value of c) against our chosen cutoff (figure five). For example, if we choose .5 as our cutoff, we will endure a 33 percent false (demographic) match rate. If we choose about .7 our cutoff, we will endure about a 19 percent false match rate. Similarly, choosing .8 as our cutoff will force us to endure about a 17 percent false match rate. The erratic increase at the right hand side is caused by a very small number of predictions at the highest probability levels.

4.3 Conclusions

The results of this evaluation indicate that administrative records addresses and households do have potential use in the Nonresponse Followup or imputation phase of a traditional census. However, the results also suggest that some caution in the use of administrative records data is in order, and improvements in processing, record linkage, modeling, and data quality will need to be made for future use of administrative records.

4.3.1 Our link rates are low enough to hurt the use of administrative records addresses in a traditional census.

Recall that approximately 81 percent of the AREX addresses linked on a one-to-one basis with a MAF address, in the AREX test sites. A small fraction of addresses linked on a many-to-many basis: Either more than one MAFID became linked with an AREX ID, or more than one AREX ID became linked with one MAFID, or both. Currently, an estimate of the number or percent of false links does not exist. Based on the finding that NRFU and imputed housing units are less likely to be linked to AREX addresses, we conclude that this will necessarily hurt (but not preclude) the use of administrative records as an aid for NRFU substitution or imputation.

4.3.2 The AREX experiment results suggest that we need continued improvements in our computerized record linkage techniques.

The 81 percent link rate suggests that continued improvements in administrative record data cleaning and standardization, and in developing tools for address record linkage across databases, has the potential to yield *significant* benefits in increasing linkage rates. However, without an assessment of false linkage rates and their characteristics, we are hampered in what we can say about the overall success at linking addresses, and hence matching household demographics.

4.3.3 Overall, for linked households, we match numbers of occupants reasonably well.

When we compare basic household demographics between AREX households and Census households, we saw that in approximately 51 percent of the linked households (52 percent of the linked occupied addresses), the AREX household count was the same as the Census household count. In almost 80 percent of the linked households, the AREX and Census household counts were the same. This suggests that administrative records are successfully predicting *how many* persons are in these addresses.

4.3.4 Overall, for linked households of the same size, we match age, race, sex, and Hispanic origin relatively well.

In about 80 percent of linked occupied households, AREX and Census agreed in demographic composition. The agreement rate is lower when we require that AREX and Census agree in both size and demographic composition. In 42 percent of all linked occupied households, AREX and Census agreed in both size and demographic composition. The numbers were not as good for NRFU households. Among linked occupied NRFU households of the same size, 63.4 percent agreed in both size and demographic composition. Among linked, occupied households in NRFU, 24 percent agreed in both size and demographic composition.

4.3.5 The race imputation model apparently is the primary cause of difficulties matching age, race, sex, and Hispanic origin within linked households.

When comparing detailed household demographics between AREX households and Census households, we find that the AREX race imputation models created substantial within-household demographic matching problems. Not only do the race imputation models create too many multi-race households, but they do so in an independent probabilistic manner, essentially "scattering" persons among different addresses. Overall, addresses where no person had their race imputed were twice as likely to match demographics with the census address than those for which at least one person had an imputed race.

4.3.6 We can predict, to a modest extent, which households are prime candidates for substitution.

We developed a logistic regression model that predicts when an AREX address will match Census age, race, sex, and Hispanic origin. If we wish to equalize false positive predictions with false negative predictions, we correctly predict match status 72 percent of the time. If we choose a more stringent cutoff, for example, requiring the predicted probability of a match to be 80 percent or greater, we correctly predict match status only 64.4 percent of the time. However, with this stricter cutoff, we successfully identify 98.8 percent of the matches, with a false positive rate of 16.8 percent.

Factors that predict demographic matches include: one or two person households, households with exclusively older persons, households where members are captured by more than one administrative record system, households with no race imputation, and households that are single-unit structure.

5. RECOMMENDATIONS

This section contains our recommendations for future work.

5.1 Improve record linkage techniques.

The success of a Bottom-Up style administrative records census depends on the ability to link addresses. Administrative records addresses must be linked with addresses on a separate address list. About 80 percent of the Census addresses linked with an AREX address on a one-to-one basis. Had the one-to-many and many-to-one links been resolved, that link rate would have improved to as high as 85 percent. However, a significantly smaller percentage of Census NRFU and imputed households were linked with AREX households. This latter fact presents a challenge for the prospects of using administrative records to substitute for nonresponse. We noted that many of the failures to link addresses by computer were due to incorrect parsing of addresses into fields, or to failure to standardize different forms of addresses that refer to the same housing unit.

Recommendation: Research into new methods of computer linkage of records should continue. New computer programs for parsing and standardizing addresses should be developed. Typically, as in AREX, a record linkage process involves a computer match followed by a clerical review process to resolve questionable links and to find links for unmatched addresses. This clerical review process should maintain, as one of its emphases, the resolution of one-to-many and many-to-one links.

5.2 Investigate ways to reduce the time lag between administrative records and surveys or censuses.

We believe the time lag between the administrative records used in AREX and Census date was a major reason for discrepancies at the household level between AREX and Census results.

Recommendation: Ways to reduce the time lag between administrative records, and when they are available for nonresponse use should be investigated. In particular, the possibility of getting records on a flow basis, and of processing those records on a flow basis should be investigated.

5.3 Improve race and Hispanic origin imputation.

Imputation of race and Hispanic origin was a source of inaccuracies of AREX demographic data. AREX households containing a person with an imputed race had a lower rate of demographic agreement than other households. AREX and Census distributions of household race characteristics differed more when AREX households with imputed race were included, than when they were not. Of particular note is that, when AREX households with imputed races were included in the distribution, AREX had a much higher percentage of mixed race households.

Recommendation: The development of improved models and other techniques to impute race and Hispanic origin should continue. In particular, we recommend development of models which emphasize demographics within the household, and characteristics of nearby households.

5.4 Continue to explore techniques for predicting when administrative records household level data are likely to be accurate.

Suppose that the accuracy of administrative records has not been proven to be accurate enough for nonresponse substitution in *all* of a particular survey or census. Administrative records may still be accurate enough to substitute for *some types* of non-responding households in that survey or census.

Recommendation: Modeling techniques should be developed to predict addresses at which administrative records are likely to be accurate. These techniques should be evaluated by using them to predict household level data within a non-responding universe, and then tested – perhaps through a field operation.

5.5 Test the use of administrative records for substitution for nonresponse.

We believe that with the lessons learned in AREX, and with the recommendations mentioned above, improved methods for conducting an administrative records census can be developed. Improved methods would increase the feasibility of using administrative records to substitute for non-responding households. These improved methods should be tested. Future Census tests would be ideal candidates for these tests.

Recommendation: The evaluation of the accuracy of administrative records and their potential for use for nonresponse substitution should be included in future Census tests. The accuracy of administrative records should be assessed by comparison with Census test data. The ability of administrative records to cover the nonresponse universe should be assessed. The accuracy of the address linkage could be addressed through field operations. Field operations could be used to evaluate the validity of models that predict households for which administrative records are particularly accurate, by testing the models' predictions about non-responding households.

REFERENCES

Aguirre International (1995). *Public Concerns About the Use of Administrative Records*. Unpublished document available from the U.S. Census Bureau, July 12, 1995.

Alvey, Wendy and Scheuren, Fritz (1982). Background for an Administrative Record Census. *Proceedings of the Social Statistics Section*, Washington DC: American Statistical Association, 1982.

American Statistical Association (1977). "Report of the Ad Hoc Committee on Privacy and Confidentiality," *The American Statistician* , 31: 59-78.

Brackstone, G.J. (1988). "Statistical Uses of Administrative Data: Issues and Challenges," in Coombs and Singh, eds., *Statistical Uses of Administrative Data : An International Symposium*, Ottawa: Statistics Canada.

Buser, Pascal, Huang, Elizabeth, Kim, Jay K., and Marquis, Kent (1998). *1996 Community Census Administrative Records File Evaluation*. Administrative Records Memorandum Series # 17. Washington, DC: U.S. Bureau of the Census.

Bye, Barry (1997). "Administrative Record Census for 2010 Design Proposal." Washington, DC: United States Department of Commerce.

Coombs, J. W. and Singh, M.P. (1988). *Statistical Uses of Administrative Data : An International Symposium*. Ottawa: Statistics Canada.

Czajka, John L., Moreno, Lorenzo, Schirm, Allen L. (1997). "On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population." Washington, DC: Internal Revenue Service.

Duncan, Otis Dudley, and Duncan, Beverly (1955). A methodological analysis of segregation indices. *American Sociological Review*, 20:210-217.

Edmonston, Barry and Schultze, Charles (1995) *Modernizing the U.S. Census*, National Academy Press, Washington DC.

Farber, J.E. and Leggieri, C.A. (2002). *Building and Validating a National Administrative Records Database for the United States.* Paper presented at the New Zealand Conference on Data Integration, January, 2002.

Gellman, Robert (1997). *Report on the Census Bureau Privacy Panel Discussion*. Unpublished document available from the U.S. Census Bureau, June 20, 1997.

Huang, Elizabeth, and Kim, Jay (2000). *One Percent Sample Study Report*. Administrative Records Research Memorandum Series #42, U.S. Census Bureau.

Judson, D. H. (1992). A Graphical Method for Assessing the Goodness of Fit of Logit Models. *Stata Technical Bulletin*, *6*:17-19.

Judson, D.H., Popoff, Carole L., and Batutis, Michael (2000*). An evaluation of the accuracy of U.S. Census Bureau County Population Estimates*. *Statistics in Transition*, in press.

Knott, Joseph J. (1991). *Administrative Records*. Memorandum for Distribution List, Bureau of the Census, Washington DC, U.S. Bureau of the Census, November 12, 1991.

Myrskyla, Pekka (1991). Census by questionnaire--Census by registers and administrative records: The experience of Finland. *Journal of Official Statistics*, *7*:457-474.

Myrskyla, Pekka, Taeuber, Cynthia, and Knott, Joseph (1996). *Uses of administrative records for statistical purposes: Finland and the United States*. Unpublished document available from the U.S. Census Bureau.

Pistiner, Arona, and Shaw, Kevin A. (2000). *Program Master Plan for the Census 2000 Administrative Records Experiment (AREX 2000)*. Administrative Records Research Memorandum Series #49. U.S. Census Bureau.

Plane, David A., and Rogerson, Peter A. (1994). *The Geographical Analysis of Population: With Applications to Planning and Business*. New York, NY: John Wiley and Sons.

Prevost, Ron (1997). *The Usefulness of IRS Information Returns in the Development of a National Administrative Records Database*. Administrative Records Research Memorandum Series #12, U.S. Census Bureau.

Sailer, Peter, Weber, Michael, Yau, E. (1993). *How Well Can IRS Count the Population?* Proceedings, Government Statistics Section, American Statistical Association. Alexandria, VA: American Statistical Association.

Singer, Eleanor, and Miller, Esther (1992). R*eactions to the Use of Administrative Records: Results of Focus Group Discussions*. Census Bureau report, Center for Survey Methods Research, August 24, 1992.

Sweet, Elizabeth (1997). *Using Administrative Record Persons in the 1996 Community Census*. Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Taueber, Cynthia, Lane, Julia, and Stevens, David (2000). *The Why, What, and How of Converting Program Records and Summarized Survey Data to State and Community Information Systems*. Paper presented at the Conference, Developing Public Policy Applications with Summarized Survey Data and Community Administrative Records. Baltimore, MD, June 6-7, 2000.

Weidman, Lynn, and Alexander, Charles (1999). *Estimation for the American Community Survey: Ongoing Work, Planned Work, and Issues*. Paper presented to the Census Advisory Committee of Professional Associations Meeting, October 21-22, 1999.

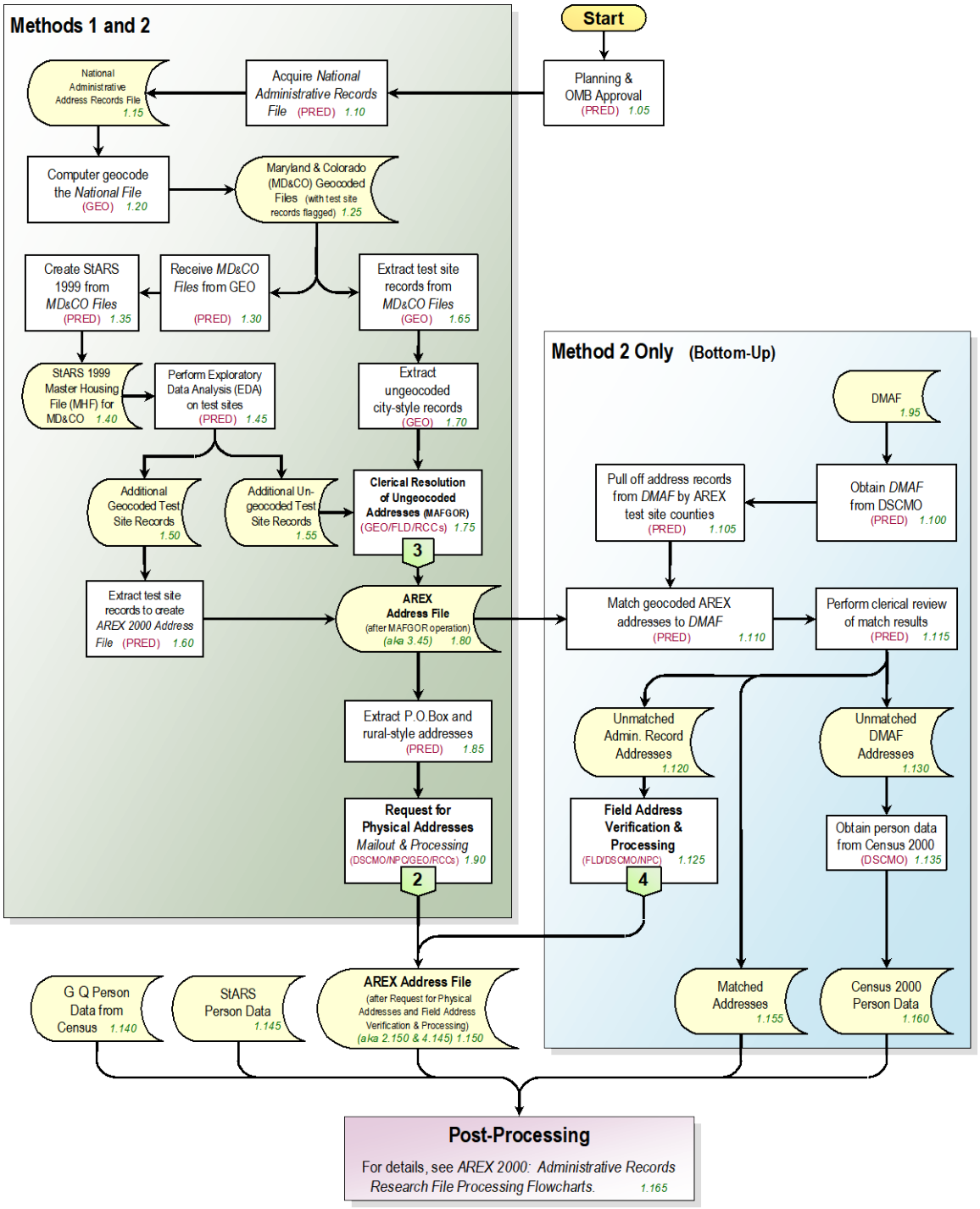
Zanutto, E. (1996). *Estimating A Population Roster rfom an Incomplete Census Using Mailback Questionnaires, Administrative Records, and Sampled Nonresponse Followup*. Presentation to the U.S. Bureau of the Census, 8/26/96.

Zanutto, Elaine, and Zaslavsky, Alan M. (1996). *Estimating a Population Roster from an Incomplete Census Using Mailback Questionnaires, Administrative Records, and Sampled Nonresponse Followup*. In Proceedings of the U.S. Bureau of the Census Annual Research Conference. Washington, DC: U.S. Census Bureau.

Zanutto, Elaine, and Zaslavsky, Alan M. (1996). *Modeling Census Mailback Questionnaires, Administrative Records, and Sampled Nonresponse Followup, to Impute Census Non-respondents*. In Proceedings, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Zanutto, Elaine, and Zaslavsky, Alan M. (2001). *Using Administrative Records to Impute for Nonresponse*. In R. Groves, R.J.A. Little, and J.Eltinge (Eds), Survey Nonresponse. New York: John Wiley.

Appendix A. AREX 2000 Implementation Flow Chart



Appendix B. Distribution Tables and Charts

Table B.1. Distributions of Household Size for the Whole AREX Universe

HH Size	Census		AREX		Census Linked to an AREX Household		Census Not Linked to an AREX Household		AREX Linked to a Census Household		AREX Not Linked to a Census Household	
1	276,590	27.2%	246,726	27.9%	229,282	25.5%	47,308	40.5%	231,223	26.9%	15,503	60.3%
2	331,472	32.6%	262,075	29.6%	297,038	33.0%	34,434	29.5%	256,745	29.9%	5,330	20.7%
3	171,136	16.8%	155,929	17.6%	155,179	17.2%	15,957	13.7%	153,199	17.8%	2,730	10.6%
4	142,822	14.0%	127,295	14.4%	131,685	14.6%	11,137	9.5%	126,046	14.7%	1,249	4.9%
5	60,988	6.0%	56,596	6.4%	56,003	6.2%	4,985	4.3%	56,064	6.5%	532	2.1%
6	21,655	2.1%	22,695	2.6%	19,866	2.2%	1,789	1.5%	22,500	2.6%	195	0.8%
7-9	11,275	1.1%	12,481	1.4%	10,335	1.1%	940	0.8%	12,359	1.4%	122	0.5%
10+	1,335	0.1%	1,625	0.2%	1,200	0.1%	135	0.1%	1,585	0.2%	40	0.2%
All Sizes	**1,017,273**	**100%**	**885,422**	**100%**	**900,282**	**100%**	**116,685**	**100%**	**859,721**	**100%**	**25,705**	**100%**

Table B.2. Distributions of Household Size for the Douglas County, Colorado

HH Size	Census		AREX		Census Linked to AREX Household		Census Not Linked to AREX Household		AREX Linked to Census Household		AREX Not Linked to Census Household	
1	8,130	13.3%	8,155	15.9%	6,533	12.5%	1,597	18.1%	7,615	15.2%	540	46.8%
2	20,930	34.4%	16,057	31.3%	17,613	33.8%	3,317	37.6%	15,753	31.4%	304	26.4%
3	11,691	19.2%	10,045	19.6%	10,052	19.3%	1,639	18.6%	9,909	19.8%	136	11.8%
4	13,277	21.8%	11,023	21.5%	11,774	22.6%	1,503	17.0%	10,911	21.8%	112	9.7%
5	5,046	8.3%	4,258	8.3%	4,486	8.6%	560	6.3%	4,213	8.4%	45	3.9%
6	1,354	2.2%	1,290	2.5%	1,191	2.3%	163	1.8%	1,277	2.5%	13	1.1%
7-9	469	0.8%	399	0.8%	421	0.8%	48	0.5%	396	0.8%	3	0.3%
10+	27	0.0%	20	0.0%	27	0.1%	0	0.0%	20	0.0%	0	0.0%
All Sizes	**60,924**	**100%**	**51,247**	**100%**	**52,097**	**100%**	**8,827**	**100%**	**50,094**	**100%**	**1,153**	**100%**

Table B.3. Distributions of Household Size for El Paso County, Colorado

HH Size	Census		AREX		Census Linked to AREX Household		Census Not linked to AREX Household		AREX Linked to Census Household		AREX Not Linked to Census Household	
1	45,945	23.9%	42,688	25.2%	39,336	22.8%	6,609	33.2%	40,239	24.4%	2,449	57.5%
2	64,060	33.3%	50,971	30.1%	57,934	33.6%	6,126	30.7%	50,109	30.4%	862	20.2%
3	32,837	17.1%	29,990	17.7%	29,872	17.3%	2,965	14.9%	29,549	17.9%	441	10.4%
4	29,922	15.6%	26,533	15.7%	27,395	15.9%	2,527	12.7%	26,222	15.9%	311	7.3%
5	12,744	6.6%	12,002	7.1%	11,628	6.7%	1,116	5.6%	11,876	7.2%	126	3.0%
6	4,534	2.4%	4,739	2.8%	4,156	2.4%	378	1.9%	4,699	2.8%	40	0.9%
7-9	2,162	1.1%	2,107	1.2%	1,976	1.1%	186	0.9%	2,089	1.3%	18	0.4%
10+	205	0.1%	247	0.1%	187	0.1%	18	0.1%	236	0.1%	11	0.3%
All Sizes	**192,409**	**100%**	**169,277**	**100%**	**172,484**	**100%**	**19,925**	**100%**	**165,019**	**100%**	**4,285**	**100%**

Table B.4. Distributions of Household Size for Jefferson County, Colorado

HH Size	Census		AREX		Census Linked to AREX Household		Census Not Linked to AREX Household		AREX Linked to Census Household		AREX Not linked to Census household	
1	50528	24.5%	47685	26.1%	43254	23.0%	7274	40.0%	44815	25.1%	2870	70.6%
2	72983	35.4%	58348	32.0%	66918	35.6%	6065	33.3%	57630	32.3%	718	17.6%
3	34106	16.6%	31468	17.2%	31773	16.9%	2333	12.8%	31205	17.5%	263	6.5%
4	30823	15.0%	27974	15.3%	29259	15.6%	1564	8.6%	27831	15.6%	143	3.5%
5	11953	5.8%	11300	6.2%	11316	6.0%	637	3.5%	11245	6.3%	55	1.4%
6	3787	1.8%	4076	2.2%	3581	1.9%	206	1.1%	4064	2.3%	12	0.3%
7-9	1699	0.8%	1601	0.9%	1601	0.9%	98	0.5%	1596	0.9%	5	0.1%
10+	188	0.1%	170	0.1%	178	0.1%	10	0.1%	168	0.1%	2	0.0%
All Sizes	**206,067**	**100%**	**182,622**	**100%**	**187,880**	**100%**	**18,187**	**100%**	**178,554**	**100%**	**4,068**	**100%**

Table B.5. Distributions of Household Size for Baltimore County, Maryland

HH Size	Census		AREX		Census Linked to an AREX Household		Census Not Linked to an AREX Household		AREX Linked to a Census Household		AREX Not Linked to a Census Household	
1	81863	27.3%	75372	27.9%	72528	26.1%	9335	41.9%	72198	27.2%	3174	63.6%
2	101341	33.8%	82613	30.6%	94567	34.1%	6774	30.4%	81617	30.8%	996	20.0%
3	51299	17.1%	48498	18.0%	48318	17.4%	2981	13.4%	48046	18.1%	452	9.1%
4	40943	13.7%	38155	14.1%	38979	14.0%	1964	8.8%	37951	14.3%	204	4.1%
5	16536	5.5%	16230	6.0%	15699	5.7%	837	3.8%	16143	6.1%	87	1.7%
6	5327	1.8%	6045	2.2%	5077	1.8%	250	1.1%	6005	2.3%	40	0.8%
7-9	2361	0.8%	2920	1.1%	2238	0.8%	123	0.6%	2895	1.1%	25	0.5%
10+	207	0.1%	317	0.1%	187	0.1%	20	0.1%	306	0.1%	11	0.2%
All Sizes	**299,877**	**100%**	**270,150**	**100%**	**277,593**	**100%**	**22,284**	**100%**	**265,161**	**100%**	**4,989**	**100%**

Table B.6. Distributions of Household Size for Baltimore City, Maryland

HH Size	Census		AREX		Census Linked to an AREX Household		Census Not Linked to an AREX Household		AREX Linked to a Census Household		AREX Not Linked to a Census Household	
1	90,124	34.9%	72,826	34.3%	67,631	32.1%	22,493	47.4%	66,356	33.0%	6,470	57.6%
2	72,158	28.0%	54,086	25.5%	60,006	28.5%	12,152	25.6%	51,636	25.7%	2,450	21.8%
3	41,203	16.0%	35,928	16.9%	35,164	16.7%	6,039	12.7%	34,490	17.2%	1,438	12.8%
4	27,857	10.8%	23,610	11.1%	24,278	11.5%	3,579	7.5%	23,131	11.5%	479	4.3%
5	14,709	5.7%	12,806	6.0%	12,874	6.1%	1,835	3.9%	12,587	6.3%	219	1.9%
6	6,653	2.6%	6,545	3.1%	5,861	2.8%	792	1.7%	6,455	3.2%	90	0.8%
7-9	4,584	1.8%	5,454	2.6%	4,099	1.9%	485	1.0%	5,383	2.7%	71	0.6%
10+	708	0.3%	871	0.4%	621	0.3%	87	0.2%	855	0.4%	16	0.1%
All Sizes	**257,996**	**100%**	**212,126**	**100%**	**210,534**	**100%**	**47,462**	**100%**	**200,893**	**100%**	**11,233**	**100%**

Table B.7. Household Race Distribution Douglas County, Colorado

HH Size			Households With all Whites # of HHs	(%[2])	Households With all of Some Race Other Than White # of HHs	(%[2])	Mixed Race Households # of HHs	(%[2])	Total[1]	(%)
	Census		7,812	(96.1%)	318	(3.9%)	N/A		8,130	**(100%)**
1	AREX	(No imputed race)	7,392	(96.5%)	270	(3.5%)	N/A		7,662	**(100%)**
	AREX	(total)	7,765	(95.6%)	358	(4.4%)	N/A		8,123	**(100%)**
	Census		19,823	(94.7%)	382	(1.8%)	725	(3.5%)	20,930	**(100%)**
2	AREX	(No imputed race)	13,641	(96.4%)	157	(1.1%)	354	(2.5%)	14,152	**(100%)**
	AREX	(total)	15,183	(94.9%)	244	(1.5%)	576	(3.6%)	16,003	**(100%)**
	Census		10,840	(92.7%)	334	(2.9%)	517	(4.4%)	11,691	**(100%)**
3	AREX	(No imputed race)	4,753	(94.7%)	85	(1.7%)	181	(3.6%)	5,019	**(100%)**
	AREX	(total)	9,231	(92.6%)	208	(2.1%)	534	(5.4%)	9,973	**(100%)**
	Census		12,268	(92.4%)	393	(3.0%)	616	(4.6%)	13,277	**(100%)**
4	AREX	(No imputed race)	3,881	(95.5%)	68	(1.7%)	113	(2.8%)	4,062	**(100%)**
	AREX	(total)	10,163	(92.7%)	218	(2.0%)	578	(5.3%)	10,959	**(100%)**
	Census		4,665	(92.4%)	126	(2.5%)	255	(5.1%)	5,046	**(100%)**
5	AREX	(No imputed race)	1,220	(93.2%)	21	(1.6%)	68	(5.2%)	1,309	**(100%)**
	AREX	(total)	3,869	(91.4%)	77	(1.8%)	287	96.8%)	4,233	**(100%)**
	Census		1,234	(91.1%)	45	(3.3%)	75	(5.5%)	1,354	**(100%)**
6	AREX	(No imputed race)	282	(91.3%)	5	(1.6%)	22	(7.1%)	309	**(100%)**
	AREX	(total)	1,146	(89.5%)	30	(2.3%)	104	(8.1%)	1,280	**(100%)**
	Census		421	(84.9%)	29	(5.8%)	46	(9.3%)	496	**(100%)**
7+	AREX	(No imputed race)	57	(80.3%)	5	(7.0%)	9	(12.7%)	71	**(100%)**
	AREX	(total)	337	(82.0%)	16	(3.9%)	58	(14.1%)	411	**(100%)**

[1] Households with no people whose race was missing

[2] Percent of Total

Table B.8. Household race Distribution El Paso County, Colorado

HH Size			Households With all Whites		Households With All of Some Race Other Than White		Mixed Race Households			
			# of HHs	(%[2])	# of HHs	(%[2])	# of HHs	(%[2])	Total[1]	(%)
	Census		41,551	(90.4%)	4,394	(9.6%)	N/A		45,945	**(100%)**
1	AREX	(No imputed race)	36,000	(90.8%)	3,626	(9.2%)	N/A		39,626	**(100%)**
	AREX	(total)	38,506	(90.4%)	4,087	(9.6%)	N/A		42,593	**(100%)**
	Census		56,090	(87.6%)	3,776	(5.9%)	4,194	(6.5%)	64,060	**(100%)**
2	AREX	(No imputed race)	39,023	(90.1%)	1,988	(4.6%)	2,288	(5.3%)	43,299	**(100%)**
	AREX	(total)	44,574	(88.2%)	2,519	(5.0%)	3,423	(6.8%)	50,516	**(100%)**
	Census		26,846	(81.8%)	2,748	(8.4%)	3,243	(9.9%)	32,837	**(100%)**
3	AREX	(No imputed race)	14,367	(84.4%)	1,229	(7.2%)	1,418	(8.3%)	17,014	**(100%)**
	AREX	(total)	24,083	(82.1%)	2,013	(6.9%)	3,246	(11.1%)	29,342	**(100%)**
	Census		24,679	(82.5%)	2,303	(7.7%)	2,940	(9.8%)	29,922	**(100%)**
4	AREX	(No imputed race)	9,627	(84.9%)	790	(7.0%)	920	(8.1%)	11,337	**(100%)**
	AREX	(total)	21,374	(82.5%)	1,503	(5.8%)	3,046	(11.8%)	25,923	**(100%)**
	Census		10,239	(80.3%)	1,093	(8.6%)	1,412	(11.1%)	12,744	**(100%)**
5	AREX	(No imputed race)	3,150	(80.9%)	327	(8.4%)	419	(10.8%)	3,896	**(100%)**
	AREX	(total)	9,231	(79.1%)	721	(6.2%)	1,718	(14.7%)	11,670	**(100%)**
	Census		3,516	(77.5%)	423	(9.3%)	595	(13.1%)	4,534	**(100%)**
6	AREX	(No imputed race)	964	(76.9%)	114	(9.1%)	175	(14.0%)	1,253	**(100%)**
	AREX	(total)	3,441	(74.9%)	314	(6.8%)	840	(18.3%)	4,595	**(100%)**
	Census		1,753	(74.1%)	227	(9.6%)	387	(16.3%)	2,367	**(100%)**
7+	AREX	(No imputed race)	272	(65.4%)	51	(12.3%)	93	(22.4%)	416	**(100%)**
	AREX	(total)	1,467	(64.7%)	175	(7.7%)	624	(27.5%)	2,266	**(100%)**

[1] Households with no people whose race was missing

[2] Percent of Total

Table B.9. Household Race Distribution for Jefferson County, Colorado

HH Size			Households With all Whites		Households With all of Some Race Other Than White		Mixed Race Households			
			# of HHs	(%[2])	# of HHs	(%[2])	# of HHs	(%[2])	Total[1]	(%)
	Census		48,891	(96.8%)	1,637	(3.2%)	N/A		50,528	**(100%)**
1	AREX	(No imputed race)	43,366	(97.5%)	1,129	(2.5%)	N/A		44,495	**(100%)**
	AREX	(total)	46,038	(96.7%)	1,553	(3.3%)	N/A		47,591	**(100%)**
	Census		69,413	(95.1%)	1,247	(1.7%)	2,323	(3.2%)	72,983	**(100%)**
2	AREX	(No imputed race)	49,120	(96.9%)	403	(0.8%)	1,161	(2.3%)	50,684	**(100%)**
	AREX	(total)	55,367	(95.6%)	648	(1.1%)	1,900	(3.3%)	57,915	**(100%)**
	Census		31,577	(92.6%)	829	(2.4%)	1,700	(5.0%)	34,106	**(100%)**
3	AREX	(No imputed race)	17,749	(95.3%)	234	(1.3%)	632	(3.4%)	18,615	**(100%)**
	AREX	(total)	28,778	(93.2%)	520	(1.7%)	1,595	(5.2%)	30,893	**(100%)**
	Census		28,465	(92.3%)	837	(2.7%)	1,521	(4.9%)	30,823	**(100%)**
4	AREX	(No imputed race)	11,624	(95.3%)	162	(1.3%)	412	(3.4%)	12,198	**(100%)**
	AREX	(total)	25,470	(92.5%)	460	(1.7%)	1,591	(5.8%)	27,521	**(100%)**
	Census		10,861	(90.9%)	413	(3.5%)	679	(5.7%)	11,953	**(100%)**
5	AREX	(No imputed race)	3,540	(93.9%)	70	(1.9%)	158	(4.2%)	3,768	**(100%)**
	AREX	(total)	10,060	(91.1%)	247	(2.2%)	737	(6.7%)	11,044	**(100%)**
	Census		3,301	(87.2%)	203	(5.4%)	283	(7.5%)	3,787	**(100%)**
6	AREX	(No imputed race)	1,004	(90.5%)	34	(3.1%)	72	(6.5%)	1,110	**(100%)**
	AREX	(total)	3,459	(87.2%)	147	(3.7%)	359	(9.1%)	3,965	**(100%)**
	Census		1,514	(80.2%)	196	(10.4%)	177	(9.4%)	1,887	**(100%)**
7+	AREX	(No imputed race)	263	(83.8%)	17	(5.4%)	34	(10.8%)	314	**(100%)**
	AREX	(total)	1,263	(74.4%)	126	(7.4%)	308	(18.1%)	1,697	**(100%)**

[1] Households with no people whose race was missing

[2] Percent of Total

Table B.10. Household Race Distribution for Baltimore County, Maryland

HH Size			Households With all Whites		Households With all of Some Race Other Than White		Mixed Race Households			
			# of HHs	(%[2])	# of HHs	(%[2])	# of HHs	(%[2])	Total[1]	(%)
1	Census		65,939	(80.5%)	15,924	(19.5%)	N/A		81,863	**(100%)**
	AREX	(No imputed race)	59,332	(81.5%)	13,446	(18.5%)	N/A		72,778	**(100%)**
	AREX	(total)	61,186	(81.3%)	14,045	(18.7%)	N/A		75,231	**(100%)**
2	Census		81,275	(80.2%)	17,385	(17.2%)	2,681	(2.6%)	101,341	**(100%)**
	AREX	(No imputed race)	62,467	(83.7%)	10,270	(13.8%)	1,929	(2.6%)	74,666	**(100%)**
	AREX	(total)	67,046	(81.4%)	12,446	(15.1%)	2,876	(3.5%)	82,368	**(100%)**
3	Census		36,308	(70.8%)	12,832	(25.0%)	2,159	(4.2%)	51,299	**(100%)**
	AREX	(No imputed race)	24,236	(76.0%)	6,497	(20.4%)	1,177	(3.7%)	31,910	**(100%)**
	AREX	(total)	35,321	(73.1%)	10,285	(21.3%)	2,686	(5.6%)	48,292	**(100%)**
4	Census		29,503	(72.1%)	9,685	(23.7%)	1,755	(4.3%)	40,943	**(100%)**
	AREX	(No imputed race)	14,130	(76.6%)	3,569	(19.3%)	754	(4.1%)	18,453	**(100%)**
	AREX	(total)	28,491	(75.0%)	7,052	(18.6%)	2,470	(6.5%)	38,013	**(100%)**
5	Census		11,424	(69.1%)	4,261	(25.8%)	851	(5.1%)	16,536	**(100%)**
	AREX	(No imputed race)	4,469	(71.2%)	1,447	(23.0%)	362	(5.8%)	6,278	**(100%)**
	AREX	(total)	11,453	(70.9%)	3,354	(20.8%)	1,356	(8.4%)	16,163	**(100%)**
6	Census		3,406	(63.9%)	1,561	(29.3%)	360	(6.8%)	5,327	**(100%)**
	AREX	(No imputed race)	1,227	(64.1%)	527	(27.5%)	161	(8.4%)	1,915	**(100%)**
	AREX	(total)	3,869	(64.3%)	1,443	(24.0%)	702	(11.7%)	6,014	**(100%)**
7+	Census		1,446	(56.3%)	877	(34.2%)	245	(9.5%)	2,568	**(100%)**
	AREX	(No imputed race)	369	(49.4%)	264	(35.3%)	114	(15.3%)	747	**(100%)**
	AREX	(total)	1,635	(50.9%)	1,009	(31.4%)	571	(17.8%)	3,215	**(100%)**

[1] Households with no people whose race was missing

[2] Percent of Total

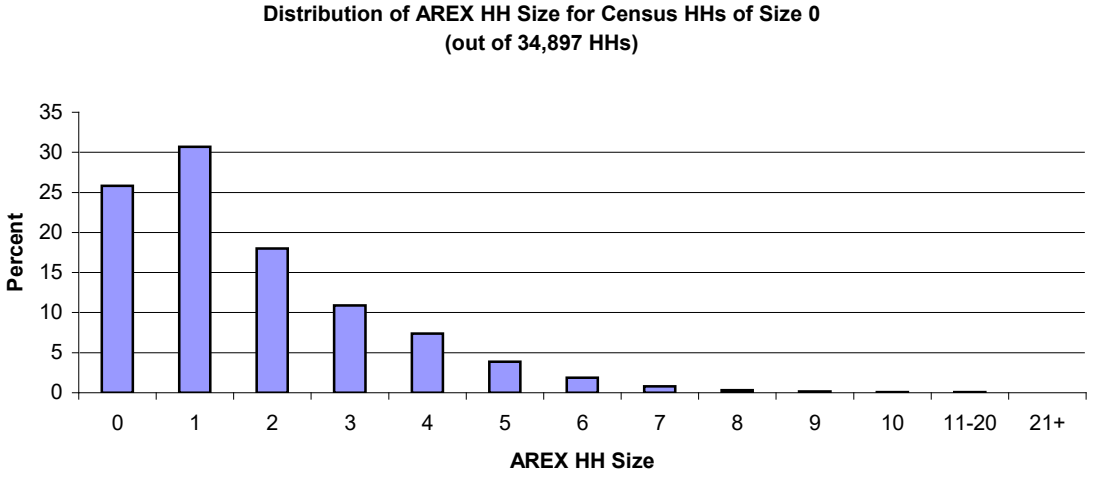
Table B.11. Household Race Distribution for Baltimore City, Maryland.

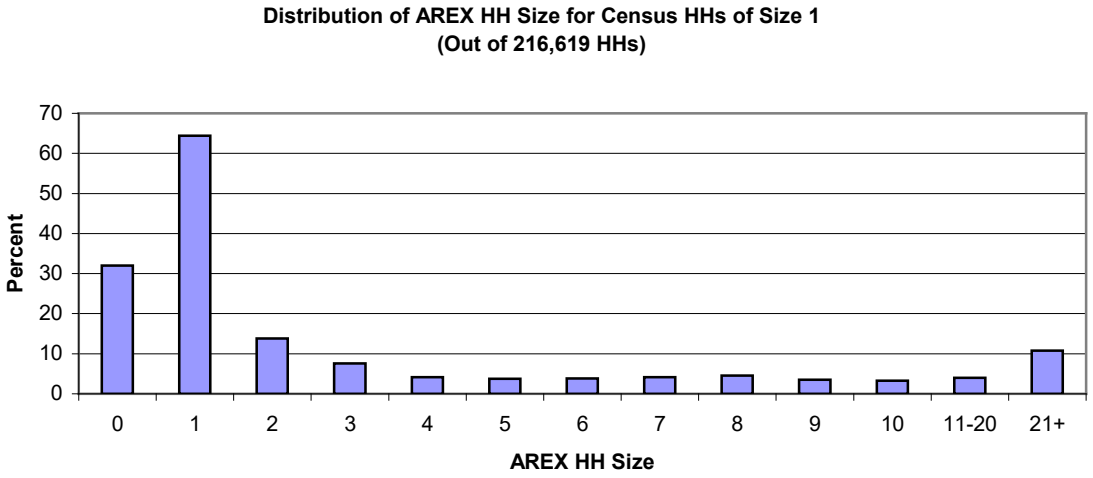
HH Size			Households With all Whites		Households With all of Some Race Other Than White		Mixed Race Households			
			# of HHs	(%[2])	# of HHs	(%[2])	# of HHs	(%[2])	Total[1]	(%)
1	Census		40,946	(45.4%)	49,178	(54.6%)	N/A		90,124	**(100%)**
	AREX	(No imputed race)	32,649	(46.3%)	37,826	(53.7%)	N/A		70,475	**(100%)**
	AREX	(total)	34,268	(47.1%)	38,434	(52.9%)	N/A		72,702	**(100%)**
2	Census		29,895	(41.4%)	39,662	(55.0%)	2,601	(3.6%)	72,158	**(100%)**
	AREX	(No imputed race)	20,977	(44.3%)	24,642	(52.0%)	1,729	(3.7%)	47,348	**(100%)**
	AREX	(total)	22,795	(42.3%)	28,521	(52.9%)	2,601	(4.8%)	53,917	**(100%)**
3	Census		11,196	(27.2%)	28,365	(68.8%)	1,642	(4.0%)	41,203	**(100%)**
	AREX	(No imputed race)	6,942	(28.1%)	16,667	(67.6%)	1,054	(4.3%)	24,663	**(100%)**
	AREX	(total)	10,061	(28.1%)	23,562	(65.8%)	2,195	(6.1%)	35,818	**(100%)**
4	Census		7,212	(25.9%)	19,382	(69.6%)	1,263	(4.5%)	27,857	**(100%)**
	AREX	(No imputed race)	3,267	(25.0%)	9,151	(69.9%)	670	(5.1%)	13,088	**(100%)**
	AREX	(total)	6,531	(27.7%)	15,166	(64.4%)	1,846	(7.8%)	23,543	**(100%)**
5	Census		3,223	(21.9%)	10,781	(73.3%)	705	(4.8%)	14,709	**(100%)**
	AREX	(No imputed race)	1,189	(19.8%)	4,412	(73.5%)	405	(6.7%)	6,006	**(100%)**
	AREX	(total)	2,920	(22.9%)	8,625	(67.5%)	1,225	(9.6%)	12,770	**(100%)**
6	Census		1,243	(18.7%)	5,037	(75.7%)	373	(5.6%)	6,653	**(100%)**
	AREX	(No imputed race)	390	(15.1%)	1,959	(76.0%)	227	(8.8%)	2,576	**(100%)**
	AREX	(total)	1,212	(18.6%)	4,599	(70.4%)	718	(11.0%)	6,529	**(100%)**
7+	Census		856	(16.2%)	4,061	(76.7%)	375	(7.1%)	5,292	**(100%)**
	AREX	(No imputed race)	175	(9.6%)	1,469	(80.4%)	184	(10.1%)	1,828	**(100%)**
	AREX	(total)	725	(11.5%)	4,685	(74.3%)	899	(14.2%)	6,309	**(100%)**

[1] Households with no people whose race was missing

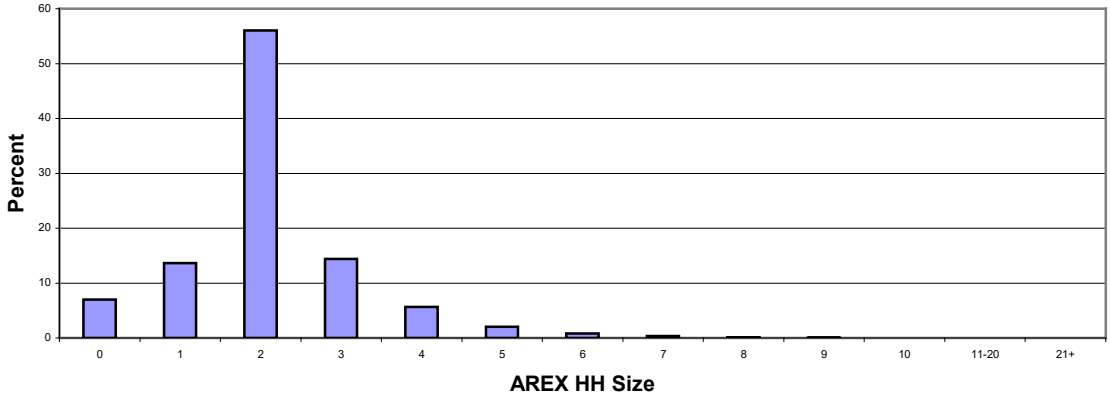
[2] Percent of Total

Charts B.12 Distributions of AREX Household Size for Fixed Census Household Sizes

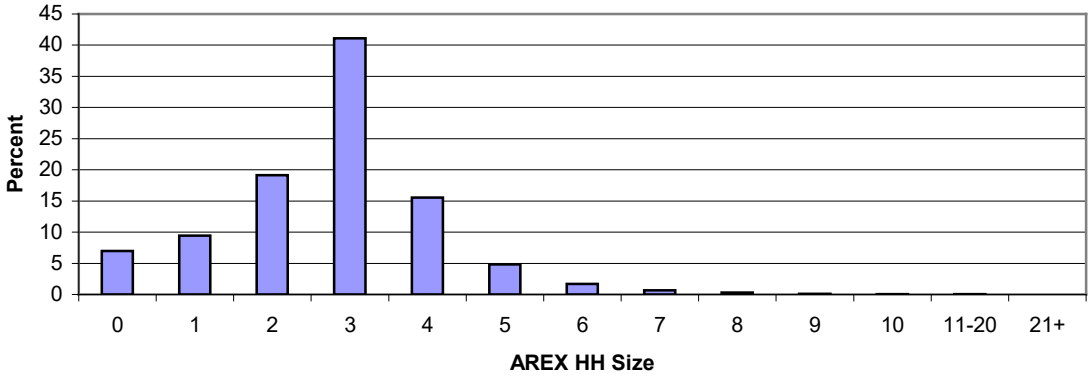




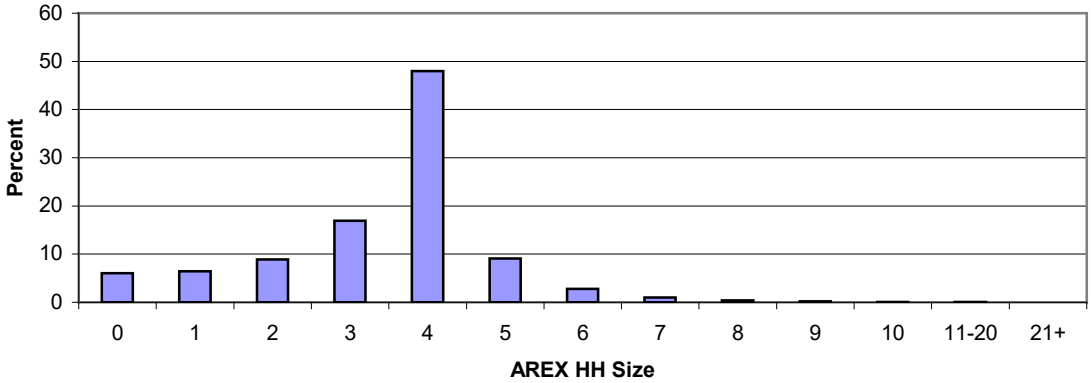
Distribution of AREX HH Size for Census HHs of Size 2
(out of 282,496 HHs)



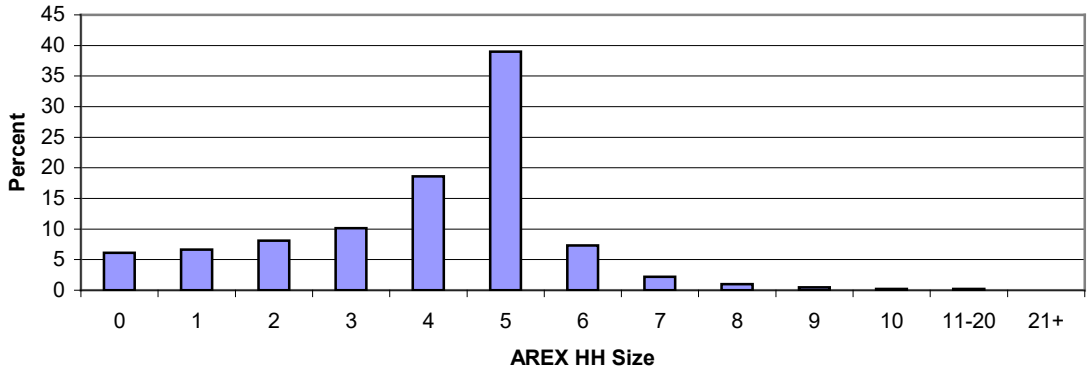
Distribution of AREX HH Size for Census HHs of Size3
(out of 147,470 HHs)



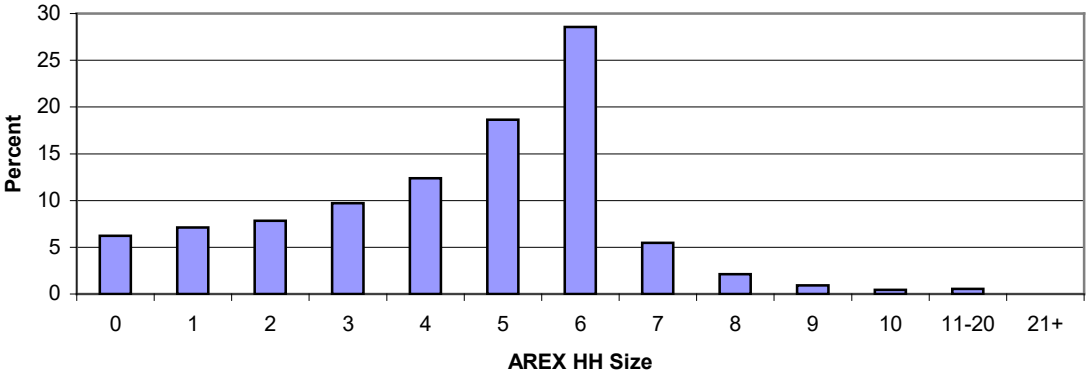
Distribution of AREX HH Size for Census HHs of Size 4
(out of 125,339 HHs)



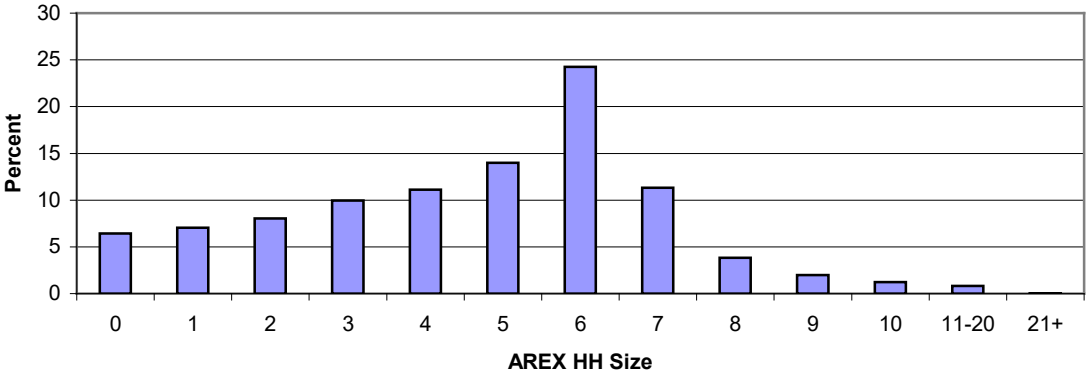
Distribution of AREX HH Size for Census HHs of Size 5
(out of 53,131 HHs)



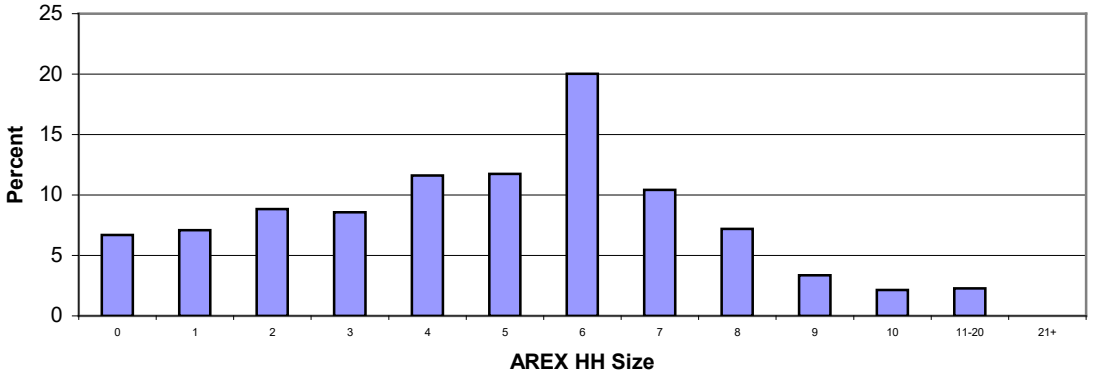
Distribution of AREX HH Size for Census HHs of Size 6
(out of 18,770 HHs)



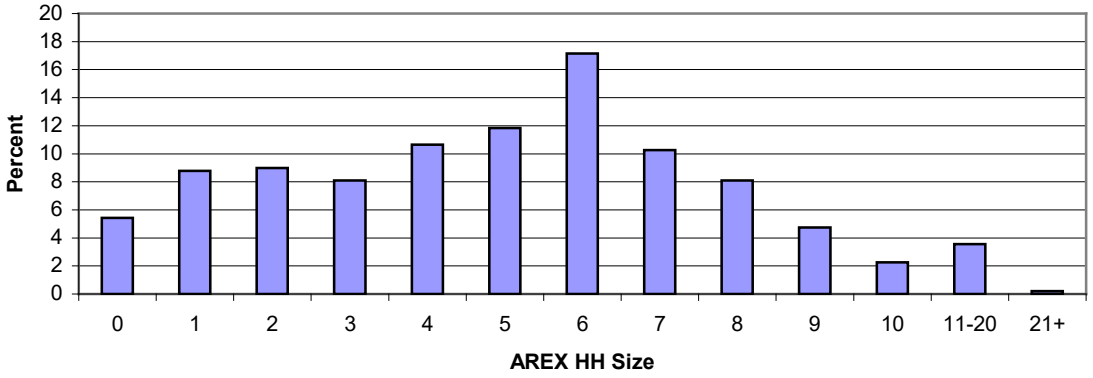
Distribution of AREX HH Size for Census HHs of Size 7
(out of 6,201 HHs)



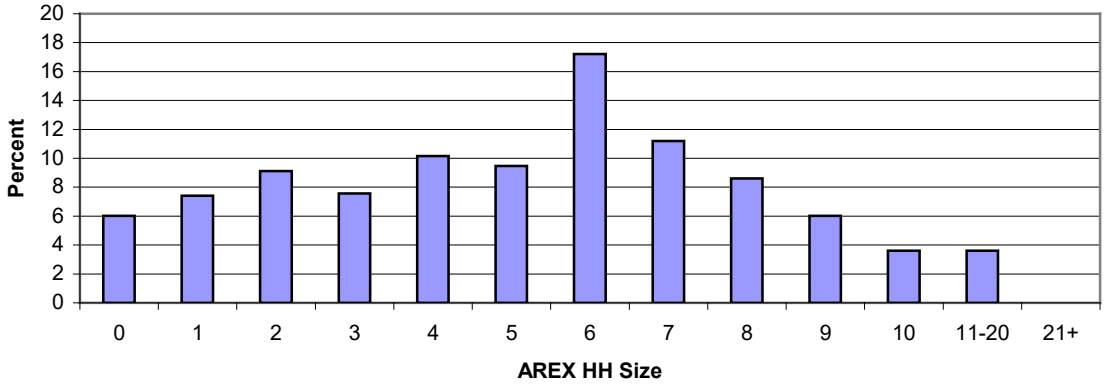
Distribution of AREX HH Size for Census HHs of Size 8
(out of 2,555 HHs)

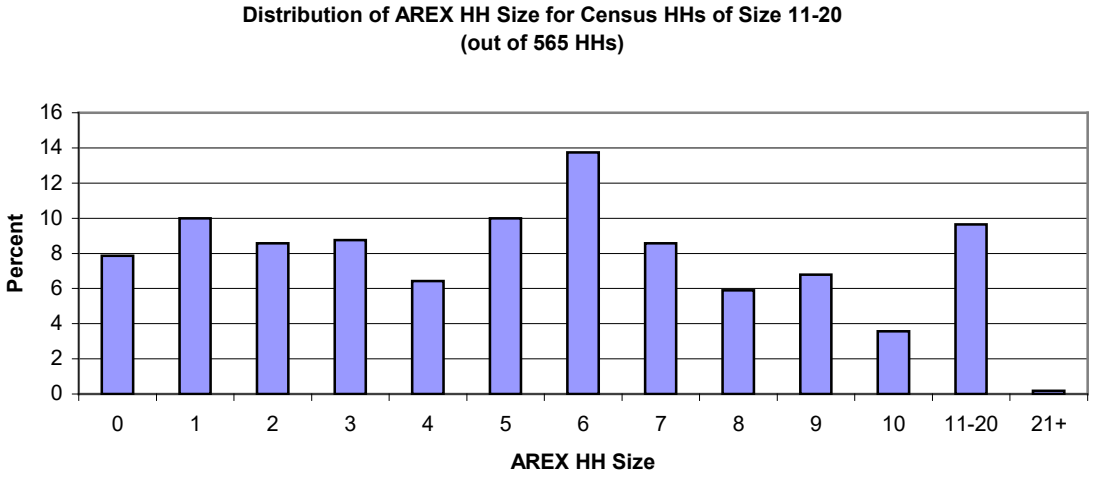


Distribution of AREX HH Size for Census HHs of Size 9
(out of 1,014 HHs)

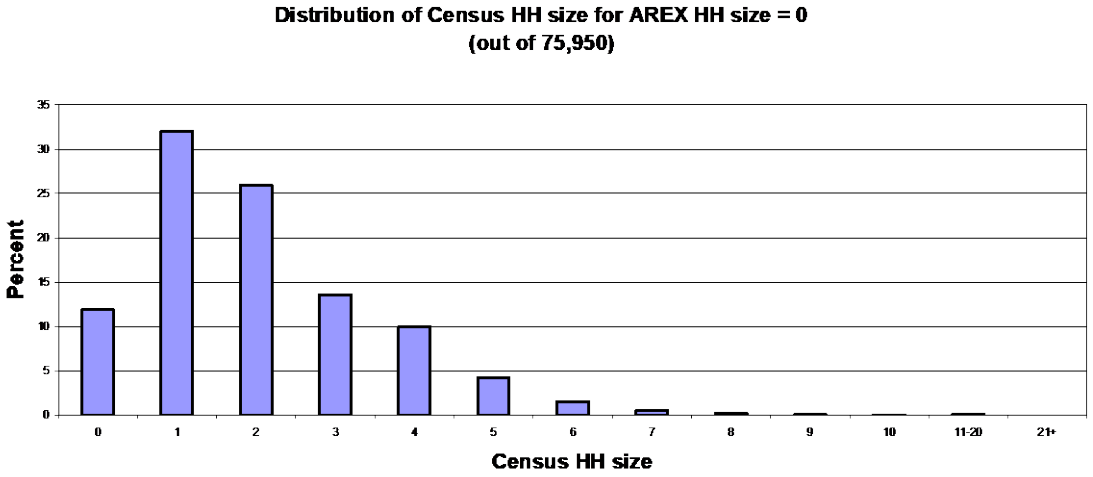


Distribution of AREX HH Size for Census HHs of Size 10
(out of 581 HHs)

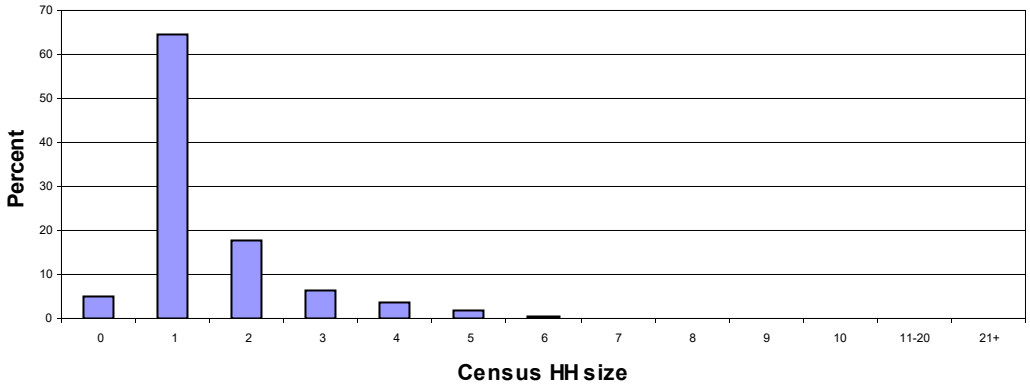



Distribution of AREX HH Size for Census HHs of Size 11-20
(out of 565 HHs)
Percent
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10
11-20
21+
AREX HH Size

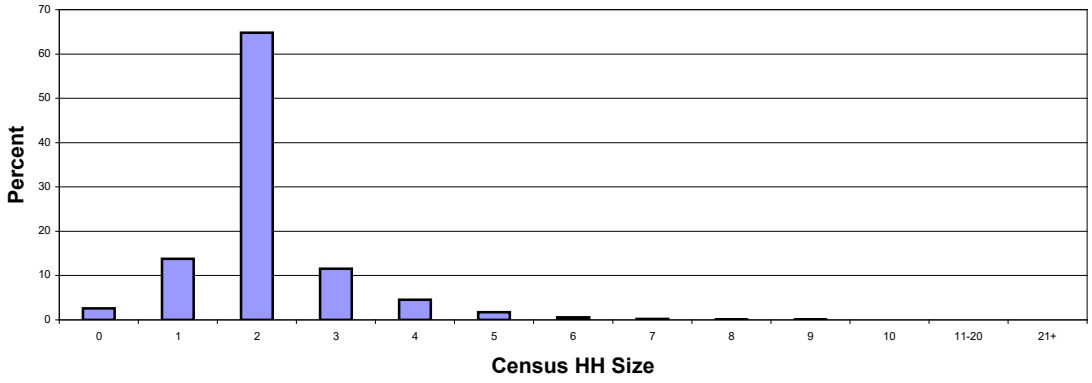
Charts B.13. Distributions of Census Household Size for Fixed AREX Household Size



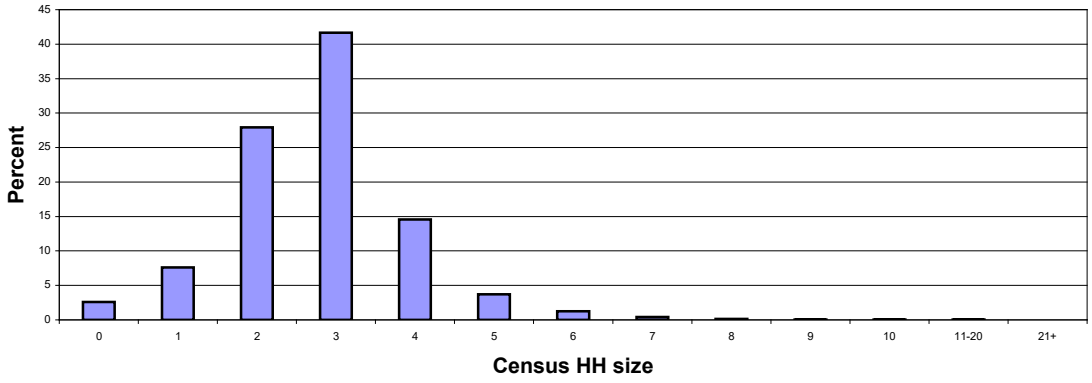
Distribution of Census HH size for AREX HH size = 1
(out of 216,209)



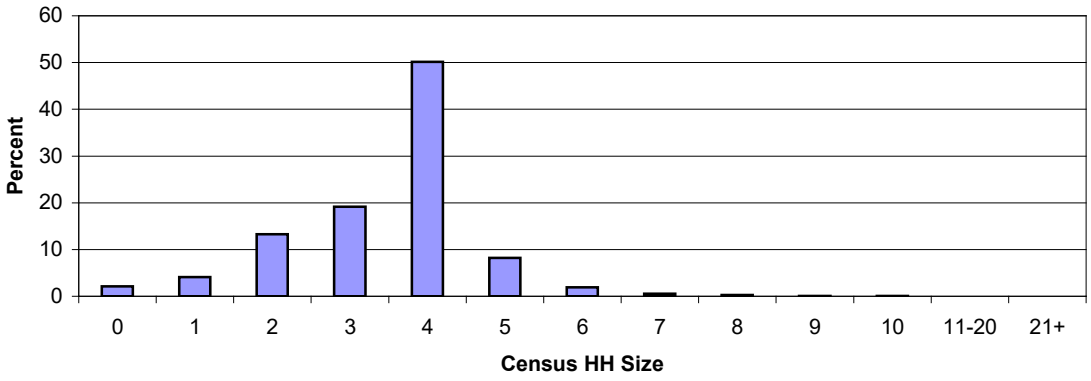
Distribution of Census HH Size for AREX HH Size = 2
(out of 244,314)



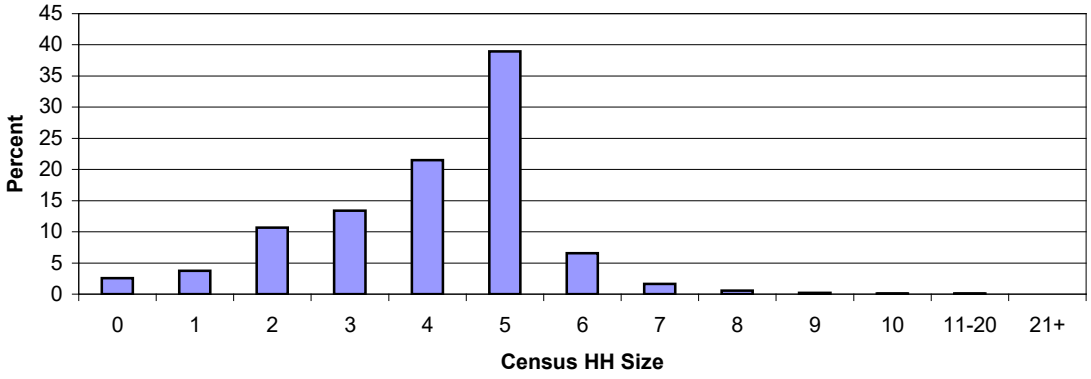
Distribution of Census HH Size for AREX HH Size = 3
(out of 145,576)



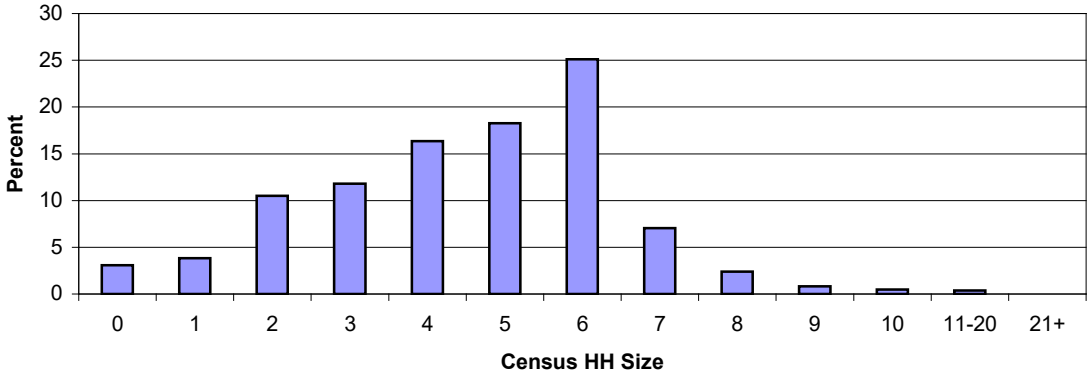
Distribution of Census HH Size for AREX HHs of Size 4
(out of 119,949 HHs)



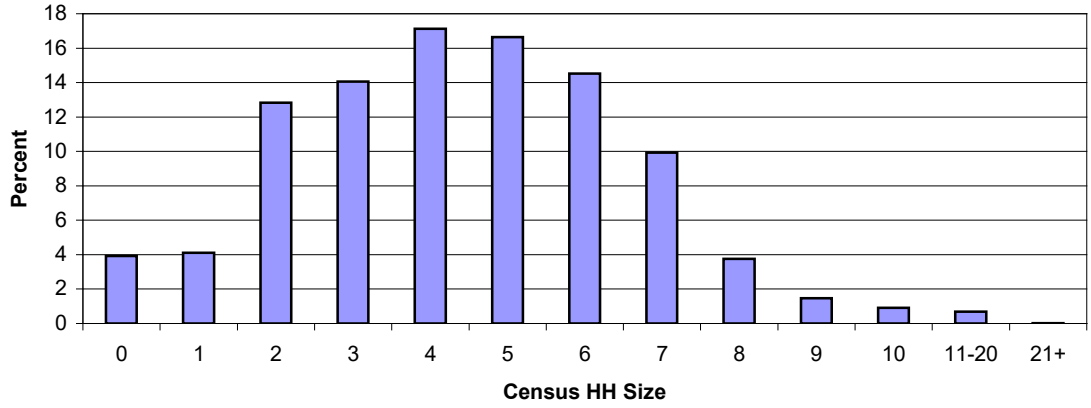
Distribution of Census HH Size for AREX HHs of Size 5
(out of 53,216 HHs)



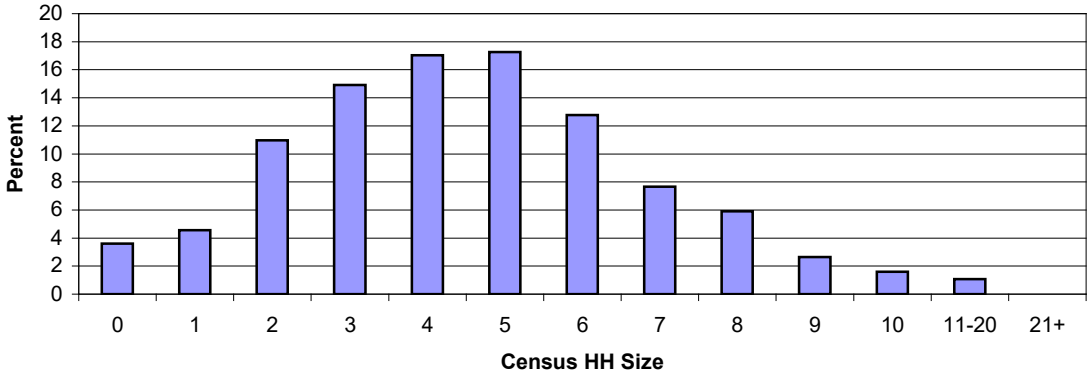
Distribution of Census HH Size for AREX HHs of Size 6
(out of 21,349 HHs)



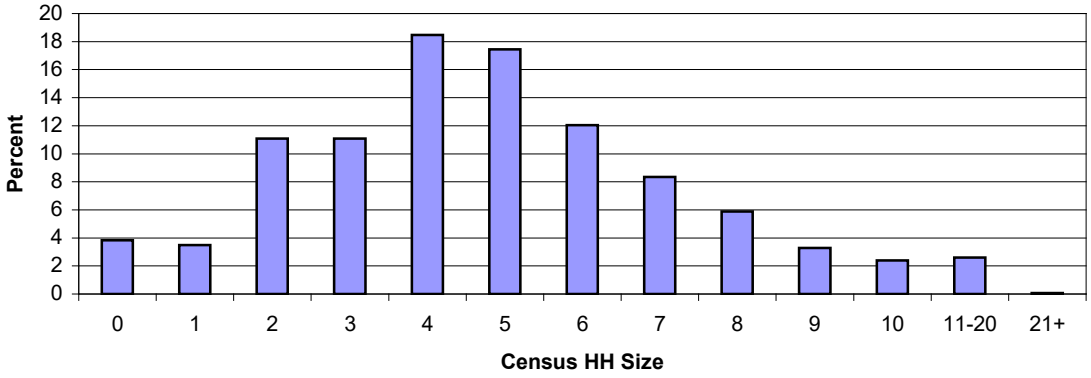
Distribution of Census HH Size for AREX HHs of Size 7
(out of 7,066 HHs)



Distribution of Census HH Size for AREX HHs of Size 8
(out of 3,110 HHs)



Distribution of Census HH Size for AREX HHs of Size 9
(out of 1,462 HHs)



Distribution of Census HH Size for AREX HHs of Size 10
(out of 699 HHs)

Distribution of Census HH Size for AREX HH Size = 11-20
(out of 701 HHs)

Distribution of Census HH Size for AREX HH Size = 21+
(out of 37 HHs)

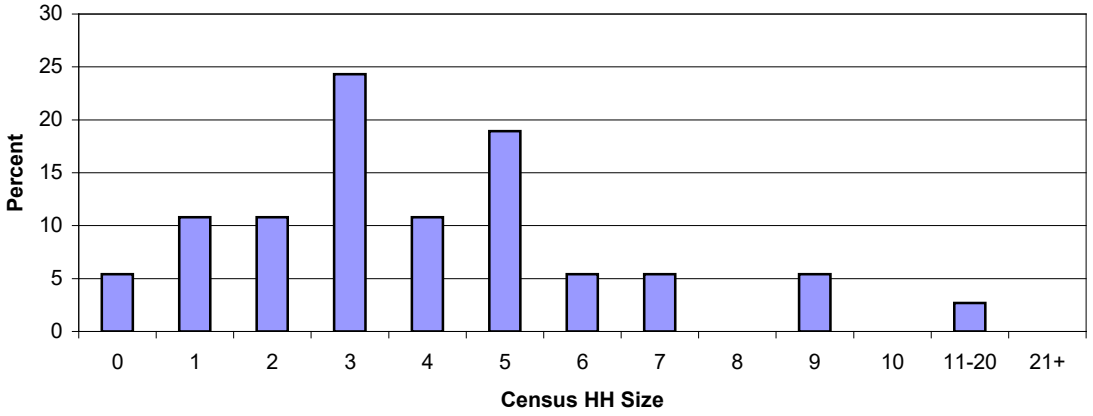


Table B.16. Coverage by AREX of Census Households by Multi vs. Single Unit, and by Household Age Characteristics

Type of Housing Unit	Census Household Age Characteristic	Total	Percent Linked
All Census HUs	All 18 or over	654,449	82.9%
	Some under 18	362,774	86.1%
Multi Unit	All 18 or over	213,722	67.6%
	Some under 18	64,725	68.7%
Single Unit	All 18 or over	440,777	90.3%
	Some under 18	298,049	89.9%
All HUs	All 50 or over	292,091	85.8%
	Some under 50	725,182	83.3%
Multi-Unit	All 50 or over	81,480	69.8%
	Some under 50	196,967	67.0%
Single-Unit	All 50 or over	210,661	91.8%
	Some under 50	528,215	89.4%
All HUs	All 65 or over	139,784	86.6%
	Some under 65	877,489	83.6%
Multi-Unit	All 65 or over	47,334	73.4%
	Some under 65	231,113	66.7%
Single-Unit	All 65 or over	92,450	93.4%
	Some under 65	646,376	89.7%

Table B.17A. AREX to Census Comparisons by Size of Housing Unit and by Household Age Characteristic

Size of Census HH	Census household age characteristic	Total	Linked with AREX Housing Units (% of Total)	Equal Size (%)1	Equal in Demographic Composition2 (%) 3
1	All 18 or over	276,490	216557	139,270	119,011
			(85.8%)	(64.3%)	(85.5%)
	Some under 18	100	62	22	1
			(62%)	(35.5%)	(4.5%)
2	All 18 or over	297,587	255,280	147,555	126,744
			(85.8%)	(57.8%)	(85.9%)
	Some under 18	33,885	27,216	10,704	6,741
			(80.3%)	(39.3%)	(63.0%)
3-4	All 18 or over	75,568	66,459	27,819	20,906
			(87.9%)	(41.9%)	(75.2%)
	Some under 18	238,390	206,350	93,003	67,459
			(86.6%)	(45.1%)	(72.5%)
5+	All 18 or over	4,854	4,041	993	568
			(83.3%)	(24.6%)	(57.2%)
	Some under 18	90,399	78,776	26,060	17,282
			(87.1%)	(33.1%)	(66.3%)

1 Percent of linked

2 Equal in: both sex groups, all four race groups, both Hispanic origin categories, and age groups 0-17, 18-64, 65+

3 Percent of linked of equal size

Table B.17B. AREX to Census Comparisons by Size of Housing Unit and by Household Age Characteristic

Type of housing unit	Census household age characteristic	Total	Linked with AREX housing units (% of Total)	Equal size (%)[1]	Equal in demographic composition[2] (%) [3]
1	All 65 or over	85,588	72,077	55,468	51,493
			(84.2%)	(77.0%)	(93.8%)
	Some under 65	191,002	144,542	83,820	67,519
			(75.7%)	(58.0%)	(80.6%)
2	All 65 or over	53,159	48,091	37,582	35,251
			(90.5%)	(78.1%)	(93.8%)
	Some under 65	278,313	234,405	120,677	98,234
			(84.2%)	(51.5%)	(81.4%)
3+	All 65 or over	1037	943	452	386
			(90.9%)	(47.9%)	(85.4%)
	Some under 65	409,971	354,683	147,423	105,829
			(86.9%)	(41.6%)	(71.8%)

[1] Percent of linked

[2] Equal in: both sex groups, all four race groups, both Hispanic origin categories, and age groups 0-17, 18-64, 65+

[3] Percent of linked of equal size